Manuel d'utilisation

STATA
Version 7.0

Valérie LAUWERS-CANCES
Pascale GROSCLAUDE
Mélanie WHITE-KONING

....Septembre 2002

Notions générales sur Stata

Stata **est** un logiciel vous permettant d'organiser vos données, de les analyser et de les représenter graphiquement. Il se compose de deux répertoires principaux sur les disquettes d'installation

- Data : qui comprend les données que vous souhaitez analyser
- Stata : qui comprend les programmes permettant la gestion des fichiers et leur analyse

Un dossier complémentaire « ado » est créé après l'installation. Il comprend toutes les commandes utilisés par le logiciel. Les mises à jour et les nouvelles commandes que vous pouvez développer vous même doivent aussi aller là.

Les mises à jour sont disponibles sur internet en saisissant la commande

➤ Update from http://www.stata.com ↵

Principes généraux

Stata fonctionne sur différents systèmes d'exploitation en particulier MAC OS, PC Window, Unix, Linux.

Les fichiers sont totalement compatibles entre Mac et PC

Stata analyse des fichiers de données auxquels il donne une structure propre. Ces fichiers peuvent provenir de différents types de fichier :

texte ascii, formaté ou avec séparateur,
fichier excel …..

Après avoir été lus, ces fichiers prennent le format Stata et peuvent être enregistrés dans ce format à votre demande. Si vous cliquez sur un fichier de données au format stata le logiciel s'ouvrira directement avec la base de données que vous venez d'ouvrir. Sinon il faudra d'abord ouvrir le logiciel et lui indiquer le chemin à suivre pour aller chercher le fichier qui vous intéresse.

Stata **n'est pas** un logiciel fait pour saisir un grand nombre de données, il ne permet ni la création d'un masque de saisie ni les contrôles lors de la saisie.

L'analyse se fait en demandant au logiciel d'exécuter des commandes Pour écrire ces commandes on utilise des (macro) commandes programmées dans le logiciel que l'on appèle en écrivant leur nom abrégé. La commande peut être précisée grâce à des options.

Le texte des commandes doit respecter une syntaxe simple propre à Stata.

Il ne faut pas utiliser de majuscules dans les commandes. Pour Stata Majuscule et minuscule d'une même lettre sont deux caractères différents

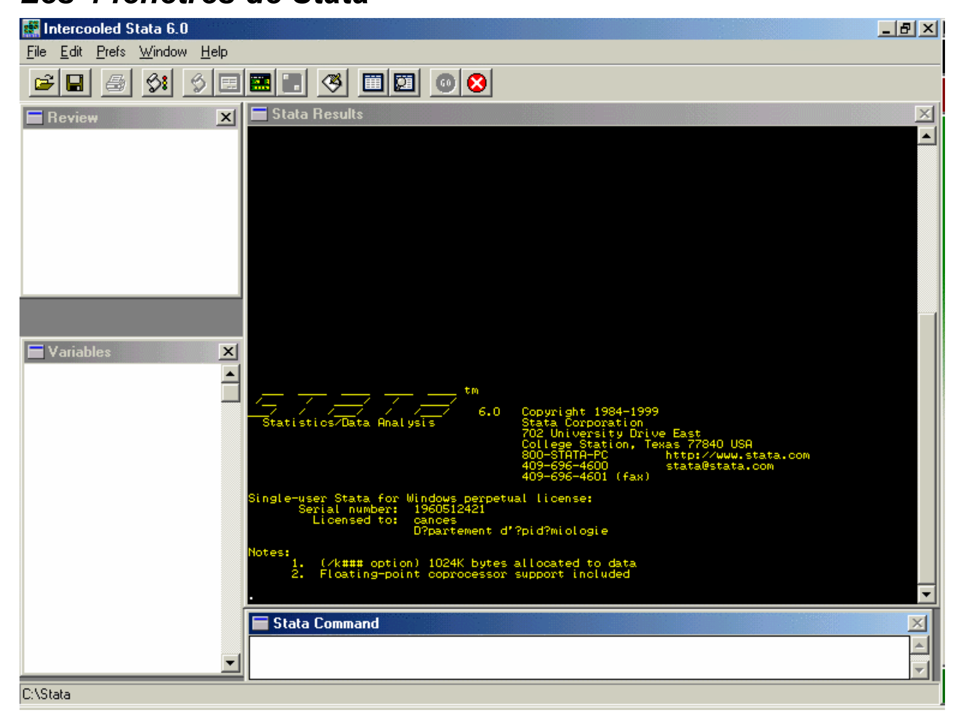
L'analyse peut se faire soit en mode interactif, soit en mode batch.

En mode interactif, l'utilisateur tape une commande, attend la réponse de la machine, tape une nouvelle commande, etc ..
En mode batch , l'utilisateur lance une séries de commandes prés enregistrées, le logiciel les exécute et rend la main à l'utilisateur à la fin de l'exécution (ou s'il s'est planté à cause d'une commande mal écrite).
Les deux modes de travail peuvent être associés dans une même session.

Extensions de fichiers utilisés pour identifier les différents types de fichiers

Mabase**.dta** :	fichier de données formatées pour stata
Monprogramme**.do** :	fichier de programme (pour le mode batch)
Mesresultats**.log** :	fichier de sortie des résultats
Mesgraph**.gph** :	fichier de sauvegarde des graphiques
Mesgraph**.wmf** :	fichier de graphique au format Windows metafile pour les PC
Mesgraph**.pic**	fichier de graphique au format Pict pour les MAC

Les 4 fenêtres de Stata



Lorsque vous ouvrez Stata 4 fenêtres apparaissent par défaut :

- Fenêtre review : permet de revenir sur les commandes effectuées depuis l'ouverture du logiciel. Ces commandes peuvent être sélectionnées par un simple click et rappelées par un double clic.
- Variables : liste des variables comprises dans la base de données. En fin de liste s'affichent les variables nouvellement crées. Ces variables peuvent être sélectionnées par un simple click
- Stata commands : zone de saisie des commandes par l'utilisateur.

- Stata Results : permet la lecture des résultats au fur et à mesure de l'analyse. Attention les résultats ne sont pas archivés automatiquement. Pour archiver les résultats il faut créer un fichier xxxx.log

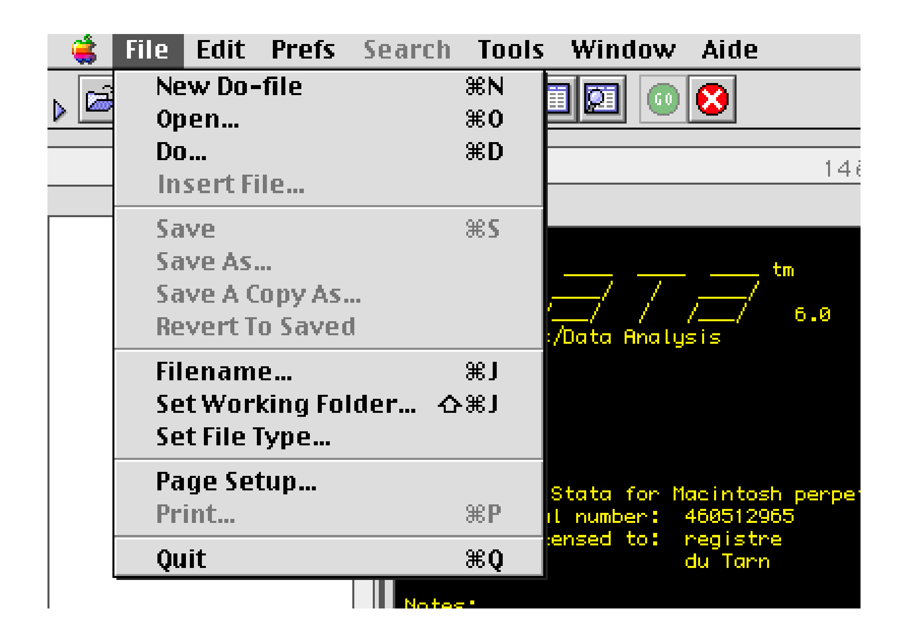
D'autres fenêtres peuvent apparaître à votre demande en cours d'analyse
Les graphiques
La base de données en cours d'utilisation
Le fichier d'archivage des résultats : xxxx.log
L'aide en ligne

Les boutons de commande et les menus



- Ouvrir : permet de choisir une base de données à analyser sous réserve que celle-ci soit enregistrée sous un format Stata ex : mabase.dta
- Enregistrer : permet d'enregistrer les modifications effectuées dans la base de données. Un message de confirmation apparaît à l'écran si vos données ont été modifiées après l'ouverture.
- Imprimer : permet d'imprimer les résultats de vos analyses
- Ouvrir un fichier de résultats : entraîne la commande log qui permet d'ouvrir un fichier de résultats qui permettra de sauvegarder les commandes que vous saisirez par la suite
- Permet de faire dérouler le fichier de résultats pour vous permettre de revenir au début de votre analyse.
- Permet de faire apparaître en premier plan et de rendre active la fenêtre de résultats
- Permet de faire apparaître en premier plan la fenêtre graphique.
- Permet de faire apparaître la base de données et de la modifier, permet de coller directement une base de données issue d'Excel.
- Permet de faire apparaître en lecture seule la base de données
- Permet de gérer le mode de défilement des résultats
- Permet d'interrompre la sortie de résultats qui est en train d'être exécutée.

D'autres commandes sont aussi disponibles dans des menus déroulant



Quelques notions concernant les variables

Dans la base de données les variables sont stockées sous des formats différents pour économiser la place occupée en mémoire.

Il ne faut pas confondre le format de stockage et la façon dont les variables s'affichent à l'écran et la présentation que vous pouvez en demander

Format d'affichage: %#.# format
% spécifie le début du format
#place occupée par la variable
.# nombre de décimaux après la virgule
format; g(général) ; f(fixe) ; e(scientifique)
Il existe aussi des format pour les variables textuelle et les dates
s(string) ; t(date)

Vous pouvez à tout moment changer le format des variables que vous utilisez pour cela vous avez besoin de savoir comment stata code les variables c'est à dire quel est le format d'enregistrement et quelle est la commande qui vous permet de passer d'un format à un autre.

Les variables qualitatives : string variables %s
Elles comprennent du texte.
Le format des données est enregistré sous forme str# # renvoyant à la place allouée à l'écriture de la variable, la longueur maximale est de 80 caractères

Les variables quantitatives : real variables

Attention si vous importez des nombres décimaux avec une « , », stata ne reconnaît pas un nombre il faut utiliser le « . ».

Ces variables sont affichées sous plusieurs formes

Format général %g	%9.0g	1.414114
Format fixé %f	%9.2f	1.41
Format scientifique %e	%9.2e	1.41^{e}+00

Variables dates %d

Elles sont obtenues après transformation des variables importées.

Elles sont stockées sous le format %t ou %d.
Stata code la date en nombre de jour écoulé à partir du 1er janvier 1960. Toutes les dates antérieures seront des nombres négatifs et toutes celles postérieures seront des nombres positifs.

Données manquantes
Elles sont codées «» pour les variables qualitatives et « . » pour les variables quantitatives. Attention le format de stockage des données manquantes pour les variables numériques est un nombre qui tend vers l'infini. Ceci implique que vous devez tenir compte des données manquantes lors de nouveaux codages ou de génération de nouvelles variables.

Les labels dans Stata

Pour ne pas avoir à interpréter des codes Stata permet de donner :
des libellés longs et plus explicites qu'un nom de 8 caractères aux variables
des libellés aux différentes modalités d'une même variable

Labeliser une variable

Label vari «ce que contient cette variable»

les guillemets sont nécessaires à cause des blancs.

Labelliser des valeurs
On commence par définir un type de label qui pourra servir pour plusieurs variables

Label define ouinon 1 oui 2 non 9 «ne sait pas»`

Puis on affecte ce label aux variables auquel il correspond

Label value ouinon diabet
Label value ouinon cardiac

Remarque : on peut aussi faire l'inverse, à partir d'une variable texte on crée une variable numérique avec la commande : encode
Encode patho , gen (maladie)

Quelques notions d'écriture des commandes (syntaxe)

[by liste de var :] commande [liste de var] [=exp] [if exp] [in range], [options]

Les annotations entre crochets sont optionnelles.
Si des variables ne sont pas précisées l'analyse portera sur l'ensemble du fichier

By var1 var2 : commande
Stipule que la commande qui suit doit être réalisée pour chaque groupe de variables
Ce préfixe ne s'utilise que lorsque les variables sont triées.

Ex

- sort sexe
- by sexe : sum age, detail

Donne la distribution de l'âge en fonction du sexe.

Commande [liste de var] [=exp]
Permet de définir le type de gestion ou d'analyse que vous effectuez
Chaque commande est suivie d'un espace

If exp
Réduit la commande à l'expression spécifiée. La commande est réalisée sur les enregistrements vérifiant l'expression logique située après if.

Ex

- sum age if sexe==2

By varlist et if peuvent être combinées dans la même ligne de commande

In range
La commande n'est appliquée qu'à une partie de l'échantillon. Celle-ci est définie par le premier numéro d'enregistrement et le dernier ## / ##

Ex

- sum age in 1/20

In range ne peut pas être combiné avec by varlist
In range peut être combiné avec if

Options

Il existe un certain nombre d'options pouvant être utilisées, elles diffèrent en fonction des commandes.
Les options sont toujours écrites après la commande principale et séparée d'elle par une virgule. Plusieurs options sont possibles au cours de la même commande.

Gérer une base de données avec Stata

Ouvrir une base stata existante

Le plus simple c'est de double-cliquer dessus mais on peut faire plus compliqué surtout si on est déjà dans stata et qu'on ne veut pas le quitter

Use
Cette commande est à utiliser quand la base de données est déjà formatée pour Stata

➢ use c:\data\mabase.dta ↵

Si vous étiez déjà en train de travailler sur un fichier, vous devez obligatoirement ajouter l'option clear

➢ use c:\data\mabase.dta, clear ↵

Clear permet de fermer la base sur laquelle vous venez de travailler.

Attention : clear ne vous propose pas d'enregistrer les modifications que vous avez apportées. Vous avez donc intérêt à sauvegarder votre fichier avant si vous y avez fait des modifications

Messages d'erreurs pouvant être rencontrés :

“ no data in memory would be lost ” : vous avez oublié de rajouter l'option clear et de sauvegarder la base que vous venez d'utiliser.

“ no room to add more records ” la base de données est trop importante, vous devez augmenter la mémoire allouée à Stata pour travailler en lui demandant avec la commande :
set mem xxm , (attention il faut enregistrer avant)

Importer des données d'une base de données extérieure

Quand la base à importer n'est pas trop volumineuse, le moyen le plus simple d'importer des données est d'effectuer un copier –coller à partir du fichier Excel. Pour cela un minimum de mise en forme est nécessaire

- Supprimer tous les accents compris dans les noms ou les champs des variables
- Donner des noms inférieurs à 8 caractères pour vos variables.
- Remplacer toutes les virgules de vos variables quantitatives par des points sinon stata les traitera comme des variables qualitatives

Puis vous copiez exactement votre fichier dans Excel, sans rajouter de colonnes ou de lignes superflues, sinon stata les traitera comme des enregistrements à données manquantes

Et enfin vous collez dans l'éditeur de données.

Insheet

Commande qui s'utilise pour importer des données en format ascii. Le fichier doit comprendre un enregistrement par ligne et des variables séparées par des tabulations ou des virgules.
Très utile

➢ Insheet using c:\data\mabase.txt ↵

Infile
Permet d'importer des fichiers en format texte dans lesquels les données ne sont pas formatées (le séparateur pouvant être une virgule, une tabulation ou un espace), les sujets pouvant être positionnés en colonne, les variables ne sont pas nommées.
Emploi peu courant, se rapporter à la documentation pour les options possibles d'importation.

➢ Infile using c:\data\mabase.raw ↵

Par défaut après infile Stata cherche un fichier .raw
Pour donner un nom aux variables lors de l'importation, vous devez les spécifier juste après la commande..

➢ Infile str10 nom age sexe using c:\data\mabase.raw ↵

Str10 sert à spécifier que la variable nom est qualitative et que la place nécessaire est de 10 caractères. Quand on ne spécifie pas le type de variables stata automatiquement en fait des variables quantitatives.

Infix
Cette commande permettant d'importer du texte en format fixe, ce qui est souvent le cas d'une base de données venant d'un autre logiciel de statistique.

Exporter des données vers une base de données extérieure

Commandes principales des bases de données

Clear : clear
Ferme la base de données en cours d'utilisation. Cette commande est nécessaire pour pouvoir lire une autre base de données

➢ clear ↵

Edit : edit
Donne accès à l'éditeur de données dans lequel il vous est possible d'effectuer des changements ou de compléter une saisie inachevée.
Je vous rappelle que Stata n'est pas un logiciel permettant les saisies importantes, celles – ci devant être réalisées sur des tableurs appropriés.

➢ edit ↵
donne accès à la base dans son intégralité

Pour ça on peut aussi utiliser le bouton, mais pas après

➢ edit var1 var2 var3↵
édite uniquement les variables listées
➢ edit var1 var2 var3 if var3<## & var2~="TT2"↵
édite les variables listées si les conditions qui suivent sont vérifiées
➢ edit in 1/20↵
édite toutes les variables des enregistrements (lignes) 1 à 20.

Save : save
Sauvegarde les données qui ont été modifiées avec les nouvelles variables créées. La commande s'écrit :

➢ save c:\data\mabase.dta↵

Si le fichier mabase existe déjà sur votre ordinateur il faut rajouter une option : "replace" qui remplace la base ancienne par la base modifiée.

Pour ça on peut aussi utiliser le menu déroulant : FILE > SAVE

➢ save c:\data\mabase.dta,replace↵

Si vous souhaitez enregistrer vos données dans un autre fichier pour garder intact le fichier d'origine, utilisez la commande save et changez le nom du fichier.
Pour ça on peut aussi utiliser le menu déroulant : FILE > SAVE AS

➢ save c:\data\mabase1.dta↵

Vous pouvez ne sauvegarder qu'un certain nombres de variables ou d'enregistrements en enrichissant les commandes de certains attributs.

➢ save nom prenom age sexe in 1/150 using c:\data\mabase1.dta↵

Sauvegarde uniquement les variables nom prénom âge et sexe des enregistrements 1 à 150.

List : list

Permet de contrôler les données dans la fenêtre résultats. Cette commande peut entraîner le défilement de l'ensemble de la base de données. Utiliser des restrictions sur les noms de variables et les enregistrements dans son utilisation.

➢ list nom prenom age sexe in 1/3 ↵

➢ list nom prenom age sexe if age<40 ↵

Si par inadvertance vous avez utilisé list vous pouvez interrompre le défilement des enregistrements par le bouton break

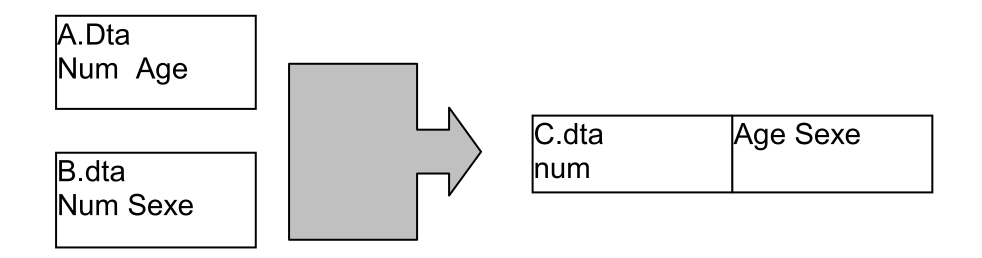
➢ list ↵
➢ Ctrl +break

Fusion de plusieurs fichiers Merge/Append

La fusion de fichiers nécessite un numéro d'identification commun aux deux fichiers et un tri préalable à la fusion sur ce numéro d'identification.

Merge

Permet de combiner les fichiers horizontalement et de rajouter des variables à la base de données. Merge ne nécessite pas que les bases contiennent le même nombre d'enregistrements.



- use c:\data\ A.dta ↵
- sort ↵
- save c:\data\A.dta, replace↵
- use c:\data\ B.dta, clear ↵
- sort ↵
- merge num using c:\data\A.dta ↵
- save c:\data\C.dta↵

Lors de la fusion stata crée une variable _merge qui vous permet de contrôler le bon déroulement de l'opération.

- tab _merge↵

```
. tab _merge
     _merge |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         18       13.85       13.85
          2 |         14       10.77       24.62
          3 |         98       75.38      100.00
------------+-----------------------------------
      Total |        130      100.00
```

Les codes s'interprètent de la façon suivante :

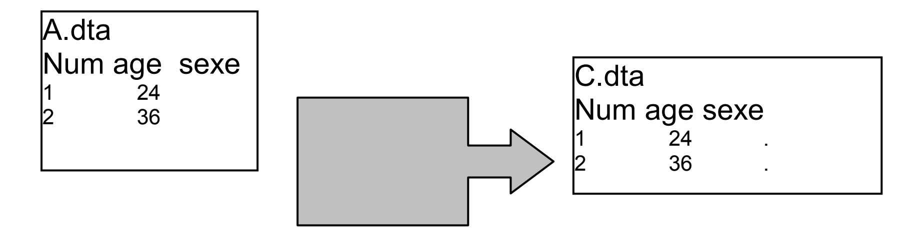
1 18 enregistrements du fichier A n'ont pas trouvé de correspondants dans le fichier B

2 14 enregistrements du fichier B n'ont pas fusionné avec le fichier A

3 98 enregistrements ont correctement fusionné entre A et B

Append

Fusion verticale des fichiers, permet d'obtenir plus d'enregistrements.





L'ordre des variables n'a pas d'importance, les fichiers peuvent être de conceptions différentes. Les données appartenant aux même noms de variables seront ajoutées, des valeurs manquantes seront spécifiées pour les variables n'existant pas dans les fichiers A et B.
Pour contrôler la fonction append, vous devez passer par l'éditeur de données et regarder s'il n'existe pas de décalage dans la base.

Outils de description de la base de données.

Codebook

Décrit les caractéristiques de chaque variable comprise dans la base. Le nombre d'enregistrements, le nombre de valeurs manquantes par variables, la distribution des variables quantitatives.

- codebook

  entraîne la description de l'ensemble de la base

- codebook var1 var2 var3

  entraîne l description des variables 1 à 3

Inspect

Moins performant que la commande précédente. Décrit la base par variables en nombre de données positives, négatives, nulles ou manquantes. Associe à la sortie des résultats un histogramme.

Describe

Permet d'obtenir les informations de synthèse de la base de données.

- Nombre d'enregistrements
- Nombre de variables
- Mémoire occupée
- Date de la commande
- Liste des variables avec leur nom et leur type

- desc
- desc, short

  supprime le listing des variables

```
Contains data from c:\data\autre.dta
  obs:           116
 vars:           184                          14 Nov 2000 00:56
 size:        87,928 (99.2% of memory free)
```

```
-----------------------------------------------------------------------
-----------
   1. num        byte   %8.0g
   2. agec1      str2   %9s                      1 agec
   3. bmij1      float  %9.0g                    1 bmij
   4. biel       float  %9.0g                    1 bmie
   5. agec2      str2   %9s                      2 agec
   6. bmij2      float  %9.0g                    2 bmij
```

Création de nouvelles variables : Gen - Egen - Replace - Recode

Gen

Cette commande crée de nouvelles variables remplissant les conditions spécifiées après le signe=
Par défaut la variable créée est toujours une variable quantitative.

➤ gen newvar = exp oldvar↵

Exp :Une combinaison de variables existantes quelles soient qualitatives et ou quantitatives

Pour les enregistrements comportant des données manquantes, la nouvelle variable prendra une valeur manquante.

Opérateurs possibles :

+	addition
-	soustraction
*	multiplication
/	division
^	exponentielle
^0.5	racine carrée
( )	regroupement d'une expression
<, <=	plus petit que
>, >=	plus grand que
~= ou !=	différent de
==	égal à
&	et
|	ou

Ex
- gen age =(dateexam-ddn) / 365.25
- gen bmi=taille/(poids^2)
- gen gold=var10>2 & (var11>1|var13>1|var15==1)

Attention lors de la définition de variable le signe = n'est pas doublé, il correspond à deux choses :
1/ l'affectation d'une valeur s'il est suivi d'une valeur,
2/ l'identification d'une expression logique s'il est suivi par une expression logique.

Par contre quand on veut parler d'égalité alors le signe est doublé ==.

Ex :

    Gen malade=diabete==1 | hta==1

Code la variable 1 si l'expression est vérifiée et 0 si elle ne l'est pas.
Malade est codé 1 si le sujet présente un diabète ou une hypertension artérielle, 0 s'il ne présente ni l'un ni l'autre

Créer des variables texte

Pour créer des variables qualitatives il faut spécifier son format et sa longueur après la commande gen et avant le nouveau nom de variable.

Gen str1 sexe= « M » if codesex==1
Replace str1 sexe= "F" if codesex==2

Plusieurs fonctions peuvent être utilisées avec les variables qualitatives

Ex : extraire les composants d'une date

- Gen str4 annee=substr(ddn,7,4)
- Gen str2 mois=substr(ddn,4,2)   fonction substr(var,#,#) *voir plus bas*
- Gen str2 jour=substr(ddn,1,2)

Convertir une variable quantitative en variable qualitative

Gen str1 genre=string(sexe)   fonction string(#)

Homogénéiser la saisie en modifiant le format (majuscule ou minuscule) des caractères

Gen str10 NOM=upper(nom)
Gen str8 prenom=lower(PRENOM)

Créer un code à partir de l'année et du mois de naissance et du sexe du patient

- Gen str7 code=sexe+annee+mois

Manipulations de chaînes de caractères plus complexes

<u>Compter le nombre de caractères dans une chaîne</u>

- gen longueur=length(nom)

```
Si le nom est "Pierre", longueur=6
```

<u>Détecter la place d'un caractère particulier dans une chaîne de caractère</u>

- gen place=index(nom,"-")

```
Si le nom est "Anne-Marie" , place=5
Si le nom est "Pierre", place=0
```

<u>Enlever des espaces</u>
ltrim : si l'espace est en début de chaîne
rtrim : si l'espace est en fin de chaîne
trim : si l'espace est en début ou en fin de chaîne

➢ gen str20 nvnom=ltrim(anciennom)
`va remplacer " Anne-Marie" par "Anne-Marie"`

Extraire les composants d'une chaîne de caractère : fonction substr(s,n1,n2)
Cette fonction soustrait une sous-chaîne de s à partir de la colonne n1 pour une longueur de n2
Si n1<0, n1 est la distance à partir de la fin de la chaîne
Si n2=. , le restant de la chaîne est renvoyé, quelle que soit sa longueur.

substr("abcdef",2,3) = "bcd"
substr("abcdef",-3,2) = "de"
substr("abcdef",2,.) = "bcdef"
substr("abcdef",-3,.) = "def"
substr("abcdef",2,0) = ""
substr("abcdef",15,2) = ""

Générer des variables dates

A partir d'une variable qualitative (string)

- gen datenais=date(ddn, "dmy", 2050)
  `remplace la variable string ddn en nombre de jours écoulés depuis le 01/01/1960`

- format datenais %d
  formate le nombre de jours en format jour/mois/année

A partir de 3 variables quantitatives jours, mois, années

Gen datenais=mdy(mois, jour, annee)

Générer une variable aléatoire

➢ gen alea=uniform()
Cette fonction renvoie des nombres pseudo-aléatoires distribués selon la loi uniforme entre 0 et 1.
Dans un second temps, cela permet de faire des tirages au sort.

Egen

Cette commande permet d'attribuer à chaque enregistrement un résumé ou une description d'une ou plusieurs variables.

Egen newvar=option(oldvar1 oldvar2 oldvarn)

Les options possibles sont :
- Count : Dénombre le nombre de valeurs non manquantes dans la liste de variables spécifiées
- diff : compare les variables, prend la valeur 1 quand les enregistrements sont différents

- max & min : donne la valeur maximale ou minimale de la variable
- mean & median : donne la moyenne ou la médiane de la variable
- pctile ## : donne la valeur du percentile spécifié
- sd : donne l'écart type de la variable
- sum : fait la somme des variables spécifiées.

Quelle est la différence entre gen et egen ?

Exemple :
gen sum1=sum(a)
egen sum2=sum(a)

list a sum1 sum2

	a	sum1	sum2
1.	1	1	15
2.	2	3	15
3.	3	6	15
4.	4	10	15
5.	5	15	15

sum1 est un résultat relatif, sum2 est une constante.

Replace

Permet de remplacer les valeurs d'une variable

- Replace poids=poids*1000
- Replace taille=taille*100

Cette commande est souvent utilisée quand l'expression logique servant à définir une variable diffère en fonction des résultats d'une autre variable.

Ex : vous voulez calculer la taille des enfants de 10 ans à 11 ans. Cette augmentation est de 10 cm pour les filles et de 8 cm pour les garçons.

- gen taille11=taille10 +8 if sexe==1 & taille10~=.
- replace taille11=taille10+10 if sexe==2 & taille10~=.

Recode

Permet de changer le codage des variables ou de définir des classes pour les variables quantitatives.

- recode csp 0=0 1=1 2=1

- gen ageclas=recode(age,20,45,75)

recode l’âge en fonction des intervalles spécifiés age<=20 sera codé 20 ; >20 et <=45 sera codé 45 ; >45 et <=75 sera codé 75

Suppression de variables ou d’enregistrements Drop - Keep

Drop

➢ drop var1 var2 var3
Supprime les variables spécifiées

➢ drop var1 var2 var3 in #/#
supprime les variables spécifiées pour un certain nombre d’enregistrement

➢ drop var1 var2 var3 if var=exp
supprime les variables spécifiées si la condition est vérifiée

➢ drop _all
supprime toutes les variables

Keep

➢ keep var1 var2 var3
Garde les variables spécifiées

➢ keep var1 var2 var3 in #/#
Garde les variables spécifiées pour un certain nombre d’enregistrements

➢ keep var1 var2 var3 if var=exp
Garde les variables spécifiées si la condition est vérifiée

Recherche et manipulation de doublons

Identification des doublons

On crée une nouvelle variable appelée **dup**

dup = 0	enregistrement unique
dup = 1	enregistrement en double, première occurrence
dup = 2	enregistrement en double, deuxième occurrence
dup = 3	enregistrement en double, troisième occurrence

etc.
La détermination des doublons se fait sur les variables **name**, **age**, et **sex**.
. sort name age sex
. quietly by name age sex: gen dup = cond(_N==1,0,_n)

_N *: nombre total d'observation dans le groupe "by"*
_n *est le numéro d'observation dans le groupe "by"*

Pour connaître le décompte des doublons , faire :

. tabulate dup

La détermination des doublons se fait sur une seule variable **name** :
. sort name
. quietly by name: gen dup = cond(_N==1,0,_n)

Comment enlever les doublons ?

Pour ne garder que la première occurrence

. drop if dup>1

Pour enlever tous les doublons
. drop if dup>0

Exemple 1
Soit les données suivantes
. list

	make	price	mpg
1.	VW Diesel	5397	41
2.	BMW 320i	9735	25
3.	Datsun 510	5079	24
4.	Audi 5000	9690	17
5.	BMW 320i	9375	25
6.	VW Diesel	5397	41
7.	BMW 320i	9735	25

Pour rechercher les doublons sur la variable **make**:

. sort make

. quietly by make: gen dup = cond(_N==1,0,_n)

Résultat :

. list

	make	price	mpg	dup
1.	Audi 5000	9690	17	0
2.	BMW 320i	9735	25	1
3.	BMW 320i	9735	25	2
4.	BMW 320i	9375	25	3
5.	Datsun 510	5079	24	0
6.	VW Diesel	5397	41	1
7.	VW Diesel	5397	41	2

. drop if dup>1
(3 observations deleted)

. list

	make	price	mpg	dup
1.	Audi 5000	9690	17	0
2.	BMW 320i	9735	25	1
3.	Datsun 510	5079	24	0
4.	VW Diesel	5397	41	1

Manipulation de fichiers

collapse

Cette commande permet de faire un nouveau fichier de données contenant des résultats (moyennes, médianes,…) provenant de variables de type quantitatif

- collapse (stat) var1 [(stat) var2], by (var3)

Pour stat; les possibilités sont les suivantes :

- Moyenne : `mean` (par défaut)
- Déviation standard : `sd`
- Somme : `sum`
- Nombre de données non manquantes : `count`
- Maximum : `max`
- Minimum : `min`
- Médiane : `median`
- Premier percentile : `p1` etc.

Exemple :

```
list marque  weight  price

        marque    weight      price
  1.    AMC        3,350      4,749
  2.    AMC        2,640      3,799
  3.    AMC        2,930      4,099
  4.    Audi       2,070      6,295
  5     Audi       2,830      9,690
  6     BMW        2,650      9,735
  7.    Buick      3,400      4,082
  8.    Buick      4,080      7,827
  9.    Buick      3,880     10,372
 10.    Buick      2,230      4,453


. collapse (mean) weight price, by(marque)

. list marque  weight  price

        marque    weight       price
  1.     AMC       2,973.3     4,215.7
  2.     Audi      2,450       7,992.5
  3.     BMW       2,650       9,735
  4.     Buick     3,397.5     6,683.5
```

contract

Cette commande permet de faire un nouveau fichier de données contenant les fréquences des combinaisons de variables

➢ contract var1 var2, options

Les options sont les suivantes :

- `freq(newvar)` : permet de donner un nouveau nom à la variable
- `zero` : pour avoir les combinaisons de variables avec des fréquences=0
- `nomiss` : les observations avec données manquantes sont éliminées

Exemple :

```
list marque  foreign rep78

        marque   foreign     rep78
  1.     AMC       Domestic        3
  2.     AMC       Domestic        3
  3.     AMC       Domestic        .
  4.     Audi      Foreign         5
  5.     Audi      Foreign         3
  6.     BMW       Foreign         4
  7.     Buick     Domestic        3
  8.     Buick     Domestic        3
  9.     Buick     Domestic        .
 10.     Buick     Domestic        3

. contract marque foreign, freq(nouveau) zero

. list  foreign marque nouveau

     foreign           marque       nouveau
  1. Domestic         AMC               3
  2. Foreign          AMC               0
  3. Domestic         Audi              0
  4. Foreign          Audi              2
  5. Domestic         BMW               0
  6. Foreign          BMW               1
  7. Domestic         Buick             4
  8. Foreign          Buick             0
```

fillin

Cette commande rajoute des observations avec des données manquantes de telle sorte que l'ensemble des combinaisons des variables spécifiées existent. Elle crée une nouvelle variable `_fillin` qui prend la valeur 1 pour les observations crées et 0 pour les observations préexistantes.

Exemple :

```
list marque foreign  rep78 price  weight

        marque   foreign     rep78      price    weight
  1.       AMC   Domestic        3      4,099     2,930
  2.       AMC   Domestic        3      4,749     3,350
  3.      Audi    Foreign        3      6,295     2,070
  4.      Audi    Foreign        5      9,690     2,830
  5.       BMW    Foreign        4      9,735     2,650

. fillin marque foreign rep78

. list marque foreign  rep78 price  weight  _fillin

        marque   foreign     rep78     price    weight   _fillin
  1.       AMC   Domestic        3     4,749     3,350         0
  2.       AMC   Domestic        3     4,099     2,930         0
  3.       AMC   Domestic        4         .         .         1
  4.       AMC   Domestic        5         .         .         1
  5.       AMC    Foreign        3         .         .         1
  6.       AMC    Foreign        4         .         .         1
  7.       AMC    Foreign        5         .         .         1
  8.      Audi   Domestic        3         .         .         1
  9.      Audi   Domestic        4         .         .         1
 10.      Audi   Domestic        5         .         .         1
 11.      Audi    Foreign        3     6,295     2,070         0
 12.      Audi    Foreign        4         .         .         1
 13.      Audi    Foreign        5     9,690     2,830         0
 14.       BMW   Domestic        3         .         .         1
 15.       BMW   Domestic        4         .         .         1
 16.       BMW   Domestic        5         .         .         1
 17.       BMW    Foreign        3         .         .         1
 18.       BMW    Foreign        4     9,735     2,650         0
 19.    BMW    Foreign    5    .    .    1
```

Quelques statistiques descriptives élémentaires

Contrôle de la nature de la distribution

SUM

➢ sum var

sum age

```
Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
   acbil |       102     9.46902    2.278692       3.75       15.2
```

Moyenne et écart type

➢ sum var, detail

```
Sum age, detail
-------------------------------------------------------------
      Percentiles      Smallest
 1%         4.25           3.75
 5%         5.75           4.25
10%         6.92           5.33       Obs                 102
25%         7.83           5.58       Sum of Wgt.         102

50%          9.2                      Mean            9.46902
                        Largest       Std. Dev.      2.278692
75%        11.08           13.5
90%        12.42          14.75       Variance       5.192439
95%         13.1             15       Skewness       .1570892
99%           15           15.2       Kurtosis       2.963282
```

Donne les estimateurs de tendances de dispersion, la nature de la distribution.
Skewness permet de juger de la symétrie
Kurtosis de l'amplitude

percentiles de la distribution

➢ pctile newvar= var, nq(10) genp(pourcent)

crée une nouvelle variable pourcent qui stipule quelle est la valeur des centiles spécifiés dans nq
Ici, il est demandé de donner les valeurs prises pour chaque décile (nq(10)).

➢ xtile newvar= var, nq(4)

crée une nouvelle variable stipulant dans quel centile se trouve la valeur de la variable quantitative. Dans cet exemple il est créé une variable à 4 classes, représentant les quartiles de la distribution.

SKTEST

➢ sktest var

. sktest acbil

Teste H0 : la distribution est normalement distribuée

```
                    Skewness/Kurtosis tests for Normality
                                                       ------- joint -------
 Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi-sq(2)  Pr(chi-sq)
----------+-----------------------------------------------------------
    acbil |       0.493          0.814            0.53        0.7661
```

SWILK

➢ swilk var

test de Shapiro – Wilk, valide pour les bases de données comportant peu d'enregistrements.

Teste l'hypothèse nulle : la distribution est normale

```
. swilk age

                   Shapiro-Wilk W test for normal data
 Variable |    Obs          W          V          z    Pr > z
----------+-----------------------------------------------------
    acbil |    102      0.99226      0.650     -0.956   0.83053
```

LADDER

Permet de tester les transformations simples permettant d'atteindre la normalité

➢ ladder var

```
ladder ins60

Transformation          formula              Chi-sq(2)      P(Chi-sq)
------------------------------------------------------------------
cube                    ins60^3                    .           0.000
square                  ins60^2                    .           0.000
raw                     ins60                  33.96           0.000
square-root             sqrt(ins60)             8.68           0.013
log                     log(ins60)              1.19           0.551
reciprocal root         1/sqrt(ins60)          18.74           0.000
reciprocal              1/ins60                39.59           0.000
reciprocal square       1/(ins60^2)            67.48           0.000
reciprocal cube         1/(ins60^3)                .           0.000
```

Tab

Permet de regarder la distribution des variables qualitatives

➢ tab var

Options pour tab :		
	row	pourcentages en ligne
	col	pourcentages en colonne
	missing	inclus les données manquantes
	chi2	chi2 de Pearson
	exact	test exact de Fisher

```
tab sexe

       sexe |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         43       37.07       37.07
          2 |         73       62.93      100.00
------------+-----------------------------------
      Total |        116      100.00
```

. tab sexe scol, row col chi2 exact

```
           |         scol
      sexe |         0          1 |     Total
-----------+----------------------+----------
         1 |        26         16 |        42
           |     61.90      38.10 |    100.00
           |     29.89      57.14 |     36.52
-----------+----------------------+----------
         2 |        61         12 |        73
           |     83.56      16.44 |    100.00
           |     70.11      42.86 |     63.48
-----------+----------------------+----------
     Total |        87         28 |       115
           |     75.65      24.35 |    100.00
           |    100.00     100.00 |    100.00

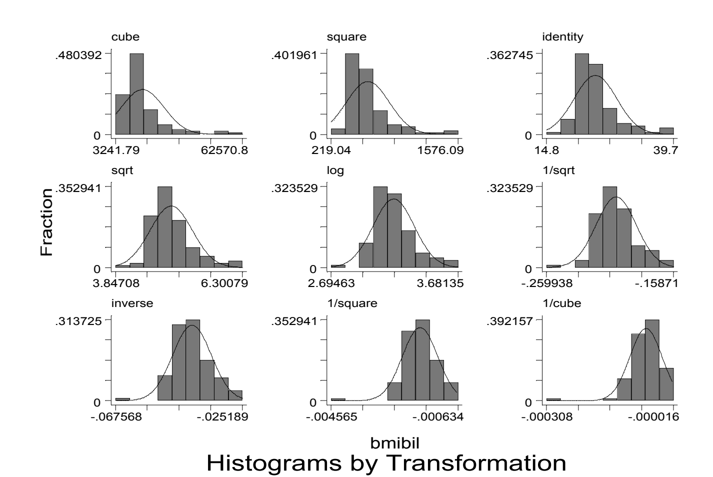
          Pearson chi2(1) =   6.7887   Pr = 0.009
           Fisher's exact =                 0.013
   1-sided Fisher's exact =                 0.009
```

Commandes Graphiques

Gladder

Visualisation graphique (histogramme) des transformations simples d'une variable quantitative.

➤ gladder var

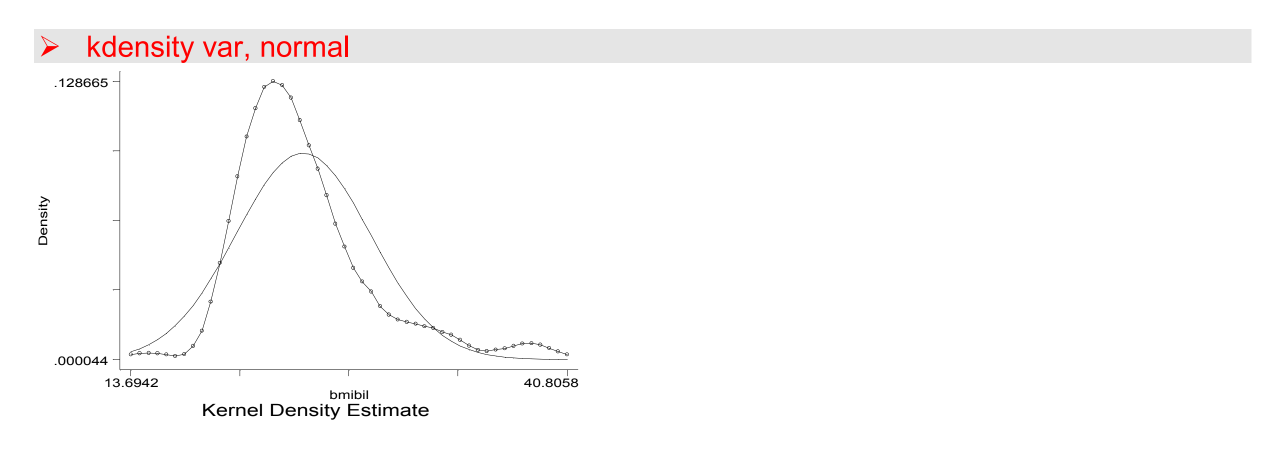


Histograms by Transformation

Kdensity

Donne la densité de f(x) : fonction de répartition de la variable. L'option normal permet le tracé d'une loi normale appliquée aux données

➢ kdensity var, normal

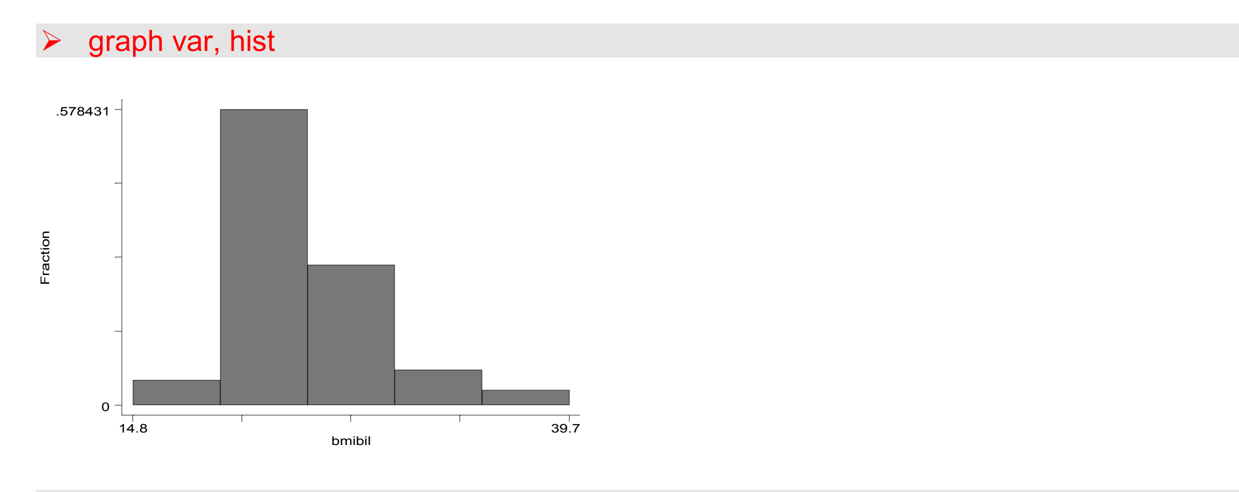


Kernel Density Estimate

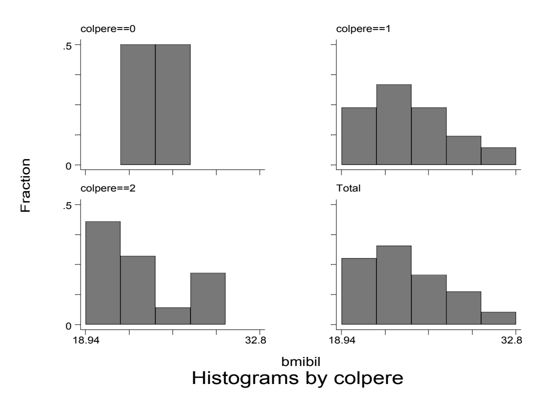
Graph

Permet de représenter graphiquement les variables
histogramme

➢ graph var, hist



➢ sort var2
➢ graph var, hist by(var2) total

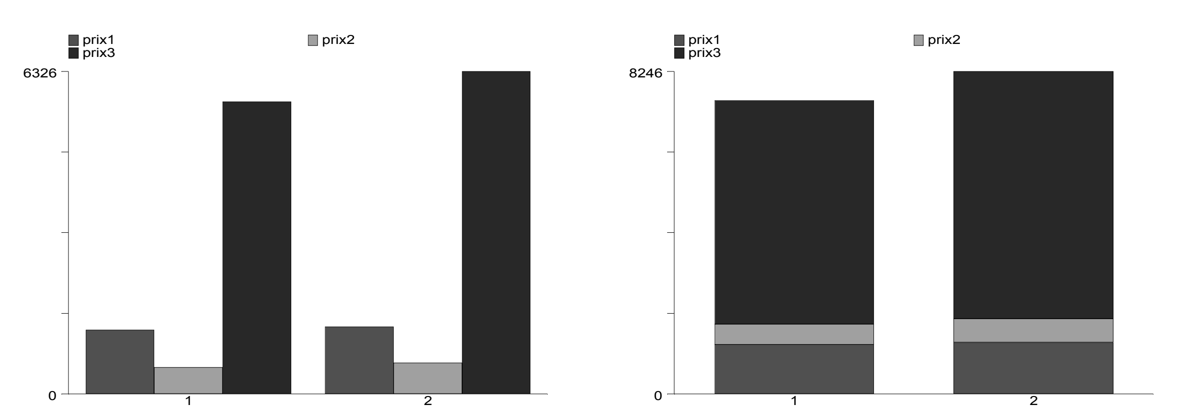


Options des histogrammes

bin (#): spécifie le nombre d'intervalles à utiliser pour les barres
freq : fréquence relative de l'axe des ordonnées
normal : dessine la distribution normale

Diagramme en barres

➢ graph var1 var2 var3, bar by(sexe) graph var1 var2 var3, bar by(sexe) stack



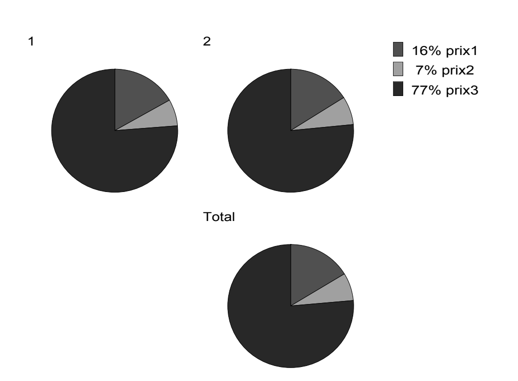
La taille des barres est proportionnelle à la somme des variables

Option des diagrammes en barre

Stack : barres empilées
Means : taille des barres est proportionnelle à la moyenne des variables
Shading :permet de gérer les couleurs des barres

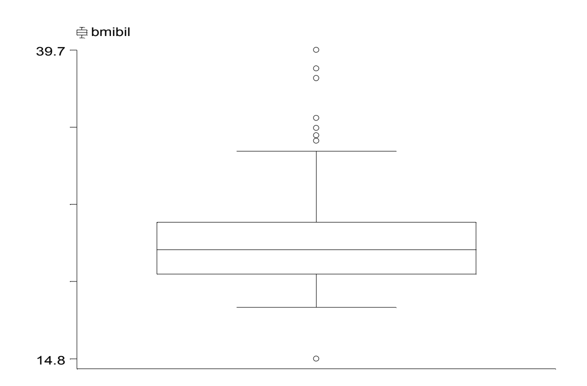
Diagrammes en secteurs

➢ graph var1 var2 var3, pie by(sexe)



box-plot

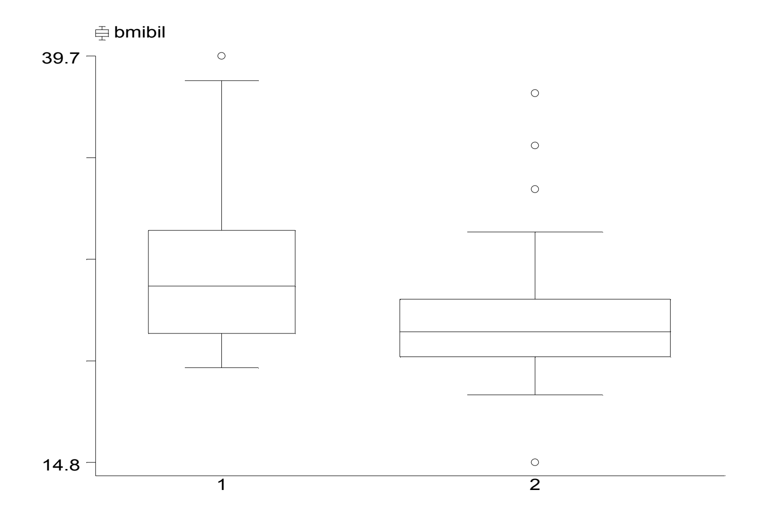
- graph var , box



Les données représentées par des points sont à vérifier, il peut s'agir de données aberrantes.
Le rectangle central correspond à la médiane de la distribution et au premier et 3ème quartile de la distribution. La distance entre le 25ème et le 75ème centile est appelée espace interquartile.
Les lignes partant du rectangle correspondent à 1,5 fois l'espace interquartile. Les données situées hors de cet espace sont à vérifier.

- sort var2
- graph var, vw box by(var2)

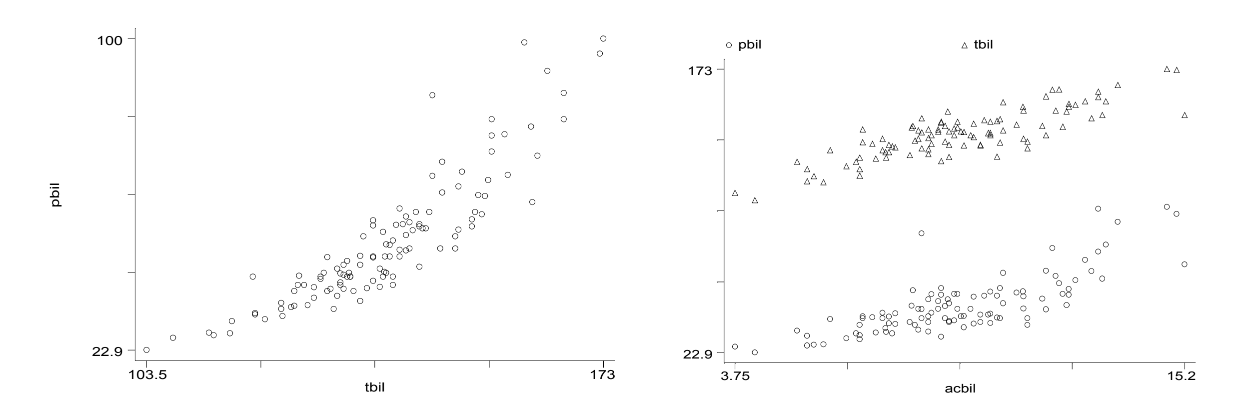
l'option **vw** proportionne les boites en fonction du nombre d'enregistrements de la variable 2



nuages de points

- graph poids taille

graph poids taille age

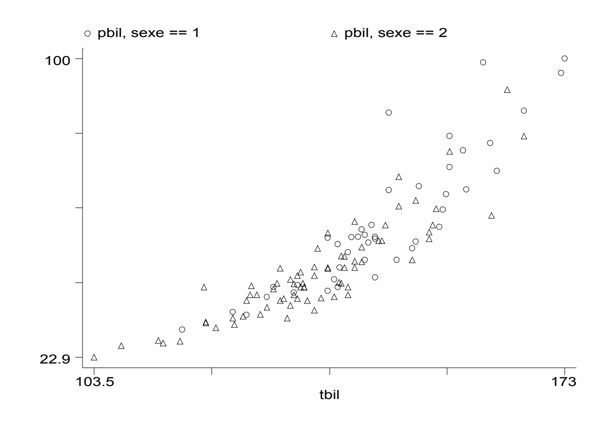


Caractérisation d'une variable qualitative (ex : sexe)

- separate poids, by(sexe)

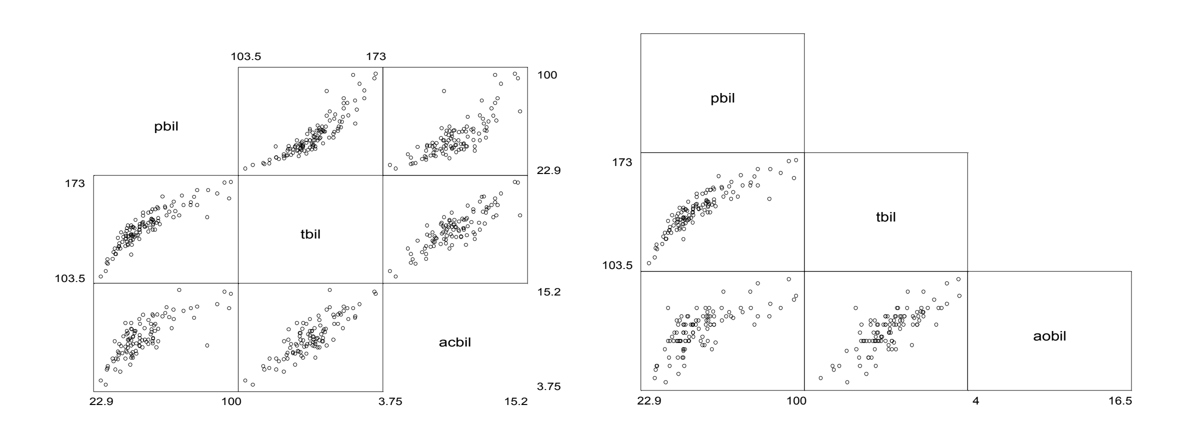
  Stata crée deux nouvelles variables nommées poids1(homme) poids2(femme)

- graph poids1 poids2 taille



Le poids des filles est marqué par un triangle
Le poids des garçons est marqué par un rond.

- graph poids taille age, matrix　　　　graph poids taille age, matrix half



options pour les matrices graphiques

- **half** donne simplement la partie inférieure de la matrice

Titre du graphique et des axes

Le titre du graphique par défaut est positionné en bas et centré. Un sous titre peut être ajouté.

Positionnement : en haut (top t), en bas (bottom b), à droite (right r), à gauche (left l)

- **title(« ___ »)** : titre centré en bas
- **t1title(« ____ »)** : titre centré en haut t2title(« ____ ») : sous-titre centré en haut
- **b1title(« ____ »)** : titre centré en bas b2title(« ____ ») : sous-titre centré en bas
- **r1title(« ____ »)** : titre centré à droite r2title(« ____ ») : sous-titre centré à droite
- **l1title(« ____ »)** : titre centré à gauche l2title(« ____ ») : sous-titre centré à gauche

options concernant les axes **:**

Par défaut, stata ne spécifie sur les axes que la première et la dernière valeur de l'étendue de la variable. Pour améliorer la présentation on peut utiliser certaines options.

- **xtick(#,…,#)** insère une marque à l'endroit spécifié sur l'axe des abscisses
- **xlabel(#,…,#)** insère une marque et le nombre correspondant sur l'abscisse
- **xscale(#)** spécifie l'étendue de l'axe des abscisses

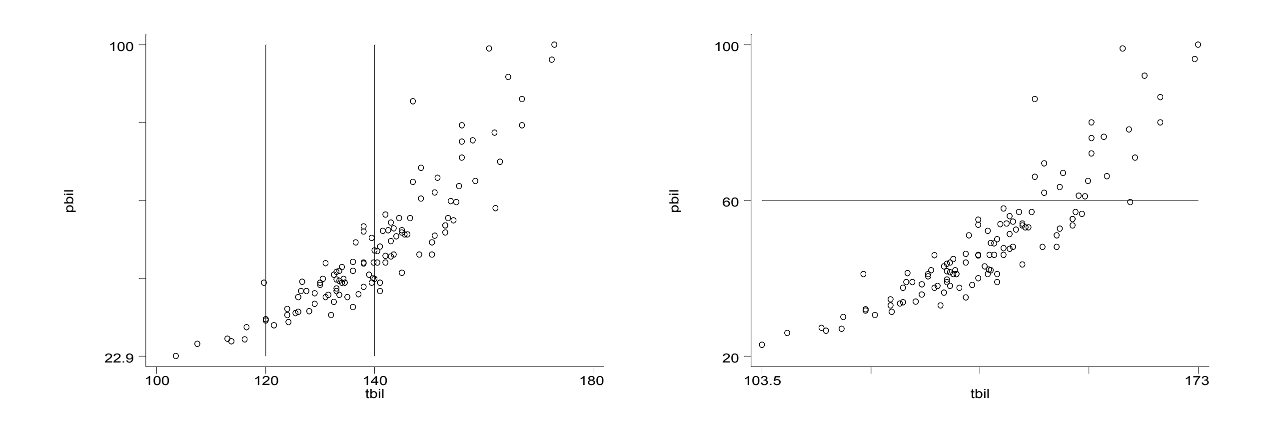
La modification de la première lettre permet de spécifier l'axe qui est concerné
x pour abscisse
y pour l'ordonnée
r pour l'axe de droite
t pour l'axe du haut

Positionnement de droites dans un graphique

Marquage d'une ou de plusieurs valeurs seuil sur les axes

- xline(#, , #)
- yline(#, , #)

➢ graph poids taille, xline(120,140) graph poids taille, yline(60)

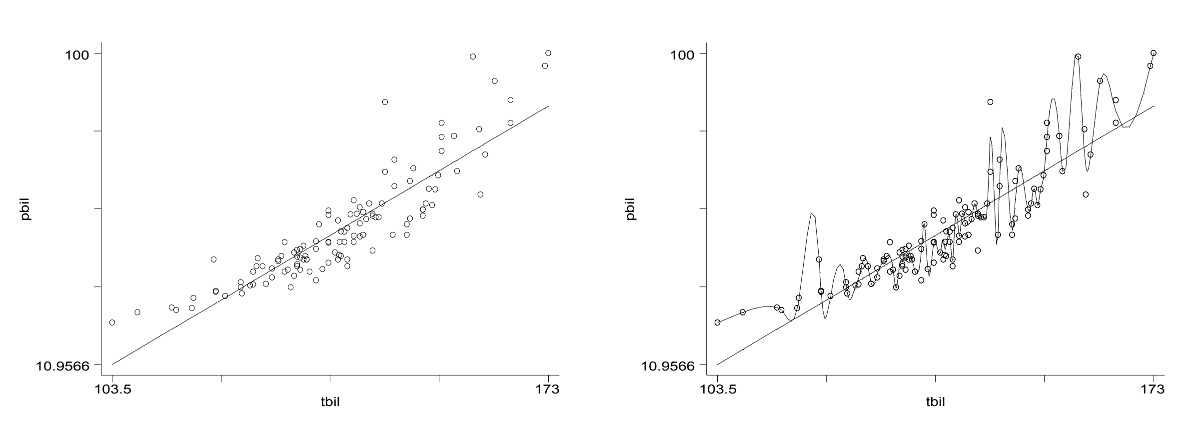


Droite de régression ou de corrélation

La droite de régression ne peut être positionnée que si elle a été au préalable estimée (regress pbil tbil / predict hat)

- graph poids hat taille , c(.l) s(Oi) sort

graph poids hat taille, c(sl) s(Oi) sort



Option Connect : relie les points entre eux

- Graph var1 var2 var3, c(.l)
    - . pas de connexion (valeur par défaut)
    - l connexion par une ligne

Option Symbol Spécifie la forme des points

- Graph var1 var2 var3, s(Oi)
    - O grand cercle / o petit cercle
    - T grand triangle
    - S grand carré
    - p plus
    - d losange
    - i invisible

s(Oi) le premier point est marqué par un cercle , le second est invisible.

Sauvegarde des graphiques

Deux formats de sauvegarde sont possibles : *.gph et *.wmf. Pour pouvoir améliorer les graphiques grâce aux outils du pack office il faut enregistrer le fichier en format windows metafile *.wmf

A partir de la barre de titre quand la fenêtre graphique est active

File
Save graph as
    Donner un nom au fichier : mongraph.wmf
    Spécifié l'option windows metafile
OK

A partir de l'option saving
Le seul format de sauvegarde possible à partir de ce fichier est un format *.gph

- graph var1 var2, saving(c:\rep\mongraph)
- graph var1 var2, saving(c:\rep\mongraph, replace)
  L'option **replace** permet d'écraser le fichier déjà enregistré

Sauvegarder plusieurs graphiques dans le même fichier

Etape-1 - Sauvegarder individuellement tous les fichiers
- graph var1 var2, saving(c:\rep\mongraph1)
- graph var1 var3, saving(c:\rep\mongraph2)
- graph var1 var4, saving(c:\rep\mongraph3)

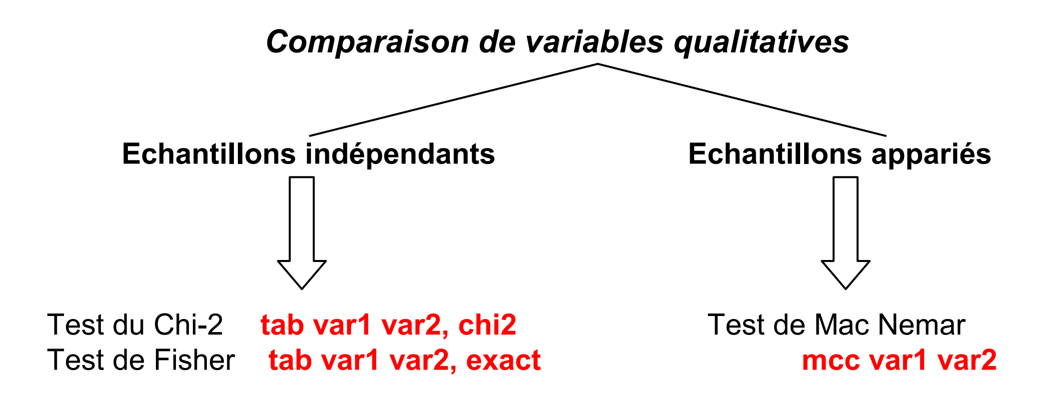
Etape-2 – Rappeler et sauvegarder tous les fichiers dans un même graphique

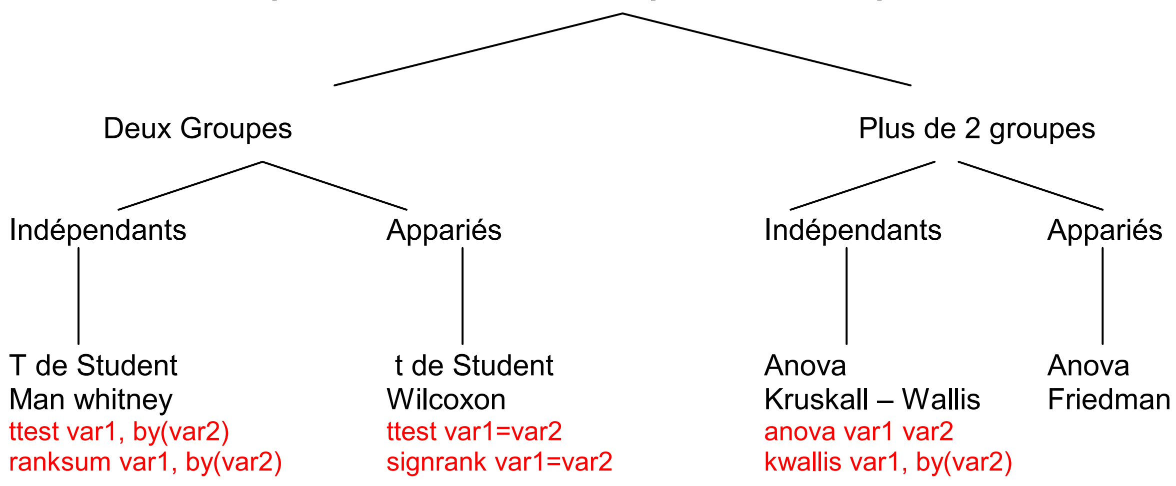
- graph using c:\rep\mongraph1 c:\rep\mongraph2 c:\rep\mongraph3, saving c:\rep\graph123

Etape-3 – Enregistrer le fichier comprenant tous les graphiques. A ce moment vous pouvez modifier le format d'enregistrement

- gphprint, saving c:\rep\graph123.wmf

Introduction à l'analyse bivariée : quels tests faut-il choisir ?







Association entre 2 variables qualitatives var1 et var2

1ère partie : séries indépendantes

1) à partir d'une base de données

➢ Tabulate var1 var2, chi2

Ex : on veut savoir s'il existe une relation entre le fait d'avoir vécu seul (variable appelée : seul) et d'être placé en maison de retraite (variable appelée institut).

L'option **chi2** (de la commande tabulate) teste l'hypothèse nulle : l'indépendance entre les 2 variables càd il y a autant de personnes ayant vécu seules que de personnes ayant vécu en couple dans les maisons de retraite. Le chi2 nous permettra de dire s'il est raisonnable de rejeter cette hypothèse.

On se fixe un seuil de décision de 5%, càd que l'on accepte de se tromper dans 5% des cas.

tabulate seul institut, row chi2

```
           |       INSTITUT
      SEUL |         0          1 |     Total
-----------+----------------------+----------
         0 |       168         36 |       204
           |     82.35      17.65 |    100.00
-----------+----------------------+----------
         1 |        38         17 |        55
           |     69.09      30.91 |    100.00
-----------+----------------------+----------
     Total |       206         53 |       259
           |     79.54      20.46 |    100.00

          Pearson chi2(1) =   4.6813   Pr = 0.030
```

On obtient un tableau à 4 cases et stata nous apprend :

- que 17.65% des personnes vivant en couple sont parties en maison de retraite alors que 30.9% des personnes seules sont parties en maison de retraite
- que le chi2 à 1 ddl est égal à 4.68, si on consulte la tables du chi2, on voit que pour 1 ddl, une valeur supérieure à 3.84 s'observe dans moins de 5% des cas.

La différence est significative, donc on observe plus de personnes ayant vécu seules en maison de retraite que de personnes ayant vécu en couple et cette différence n'est pas dûe au hasard.

Autrement dit : p = 0.03 c'est la probabilité, si l'hypothèse nulle est vraie, que le hasard explique la différence observée ici. Or, cette probabilité est inférieure à 5% donc on conclue : les différences observées ne sont pas dûes au seul hasard, il existe donc une différence significative à 5%.

On peut afficher les pourcentages en lignes et en colonnes :

tab seul institut, **row col** chi2

	INSTITUTB		
SEUL	0	1	Total
0	168	36	204
	82.35	17.65	100.00
	81.55	67.92	78.76
1	38	17	55
	69.09	30.91	100.00
	18.45	32.08	21.24
Total	206	53	259
	79.54	20.46	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 4.6813 Pr = 0.030

Remarque importante : Contrairement à certains logiciels de statistiques, stata ne fait pas de test par défaut, mais seulement ceux qu'on lui demande, même si les conditions d'application ne sont pas respectées.

Rappel : pour le test du chi2, la condition d'application est d'avoir des effectifs théoriques supérieurs à 5.

Ex : on veut savoir s'il existe une relation entre le fait d'avoir une anorexie (variable appelée : nut) et d'être placé en maison de retraite (variable appelée institut).

Test du chi2 $\chi^2 = \frac{\sum (O - C)^2}{C}$

tabulate nut institut, chi2

	INSTITUT		
NUT	0	1	Total
0	202	48	250
1	1	3	4
Total	203	51	254

Pearson chi2(1) = 7.6390 Pr = 0.006

Stata a calculé le chi2 sans donner de message d'erreur, or certains effectifs théoriques sont inférieurs à 5, et dans ce cas, le résultat du chi2 , n'est pas valide, et il est indispensable d'utiliser un test exact de Fisher.

Test exact de Fisher

➢ Tabulate var1 var2, exact

L'option **exact** (de la commande tabulate) permet d'obtenir un test exact de Fisher.

Cette option est valide pour une table 2x2 mais également pour une table r X c avec plusieurs lignes ou colonnes. Dans ce cas, le temps nécessaire à stata pour faire le calcul peut atteindre plusieurs minutes… et vous pouvez avoir recours à l'option break !

test exact de Fisher : $P = \dfrac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,(a+b+c+d)!}$

. tabulate nut institut, exact

```
           |        INSTITUT
    NUT    |         0          1 |     Total
-----------+----------------------+----------
         0 |       202         48 |       250
         1 |         1          3 |         4
-----------+----------------------+----------
     Total |       203         51 |       254

           Fisher's exact =                 0.026
   1-sided Fisher's exact =                 0.026
```

2) Calcul direct sans base de données

Ex : dans un article, l'auteur déclare qu'il existe une relation entre le fait d'avoir vécu seul et d'être placé en maison de retraite mais ne vous donne pas le résultat du test, or, vous disposez des effectifs dans un tableau décrivant sa population d'étude.

La commande **tabi** teste l'hypothèse nulle : l'indépendance entre les 2 variables càd il y a autant de personnes ayant vécu seules que de personnes ayant vécu en couple dans les maisons de retraite. Le test nous permettra de dire s'il est raisonnable de rejeter cette hypothèse.

➢ Tabi a b \ c d

Où a, b, c et d sont les effectifs du tableau à 4 cases

```
           |        col
    row    |         1          2 |     Total
-----------+----------------------+----------
         1 |         a          b |     a+b
         2 |         c          d |     c+d
-----------+----------------------+----------
     Total |       a+c        b+d |     a+b+c+d
```

Exemple : tabi 30 18 \38 14

```
           |           col
       row |         1          2 |     Total
```

```
-----------+----------------------+----------
         1 |        30         18 |        48 
         2 |        38         14 |        52 
-----------+----------------------+----------
     Total |        68         32 |       100 

           Fisher's exact =                 0.289
   1-sided Fisher's exact =                 0.179
```

NB :

- par défaut, stata donne le résultat d'un test de Fisher, mais on peut lui donner comme option chi2
- cette commande est valide pour les tableaux 2X2, mais également les tableaux r X c.

2ème *partie : séries appariées*

1) à partir d'une base de données

➢ mcc var1 var2

2) calcul direct sans base de données

Ex : dans un article, l'auteur déclare que pour une série de sujets ayant bénéficié à un mois d'intervalle de 2 traitements différents (1 et 2) qu'il existe plus d'effets indésirables avec le traitement 1 qu'avec le traitement 2 mais ne vous donne pas le résultat du test, or, vous disposez des effectifs dans un tableau décrivant sa population d'étude. NB : chaque sujet a donc reçu les deux traitements.

La commande **mcci** teste l'hypothèse nulle : il y a autant d'évènements indésirables avec le traitement 1 qu'avec le traitement 2. Le test nous permettra de dire s'il est raisonnable de rejeter cette hypothèse.

Rappel : pour le test du chi2 de Mac Nemar (ou de Mantel-haenszel), la condition d'application est d'avoir la somme des paires discordantes supérieure ou égale à 10.

➢ mcci a b c d

Où a, b, c et d sont les effectifs du tableau à 4 cases

```
           |        témoins
    cas    |   expose   non expose |      Total
-----------+-----------------------+----------
expose     |         a           b |      a+b
non expose |         c           d |      c+d
-----------+-----------------------+----------
     Total |       a+c         b+d |      a+b+c+d
```

Exemple : mcci 16 4 12 18

```
                 |     EI avec TTT1        |
EI avec TTT2     |   oui           non     |      Total
-----------------+-------------------------+----------
         oui     |         16            4 |         20
         non     |         12           18 |         30
-----------------+-------------------------+----------
           Total |         28           22 |         50

McNemar's chi2(1) =      4.00         Pr>chi2 = 0.0455
Exact McNemar significance probability        = 0.0768

Proportion with factor
        Cases              .4
        Controls          .56      [95% conf. interval]
                    ---------      --------------------
        difference       -.16      -.3303945    .0103945
        ratio        .7142857       .5128527    .9948355
        rel. diff. -.3636364      -.7797718    .0524991
        odds ratio   .3333333       .0783559    1.099907    (exact)
```

Comparaison de moyennes

Par rapport à une moyenne théorique

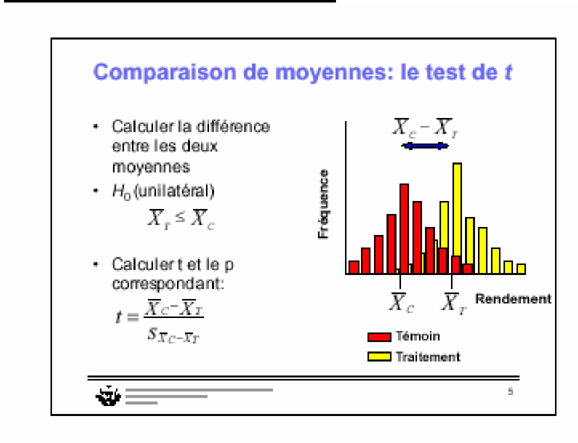
➤ ttest var1 = #↵

- `ttest age =10`

Permet de comparer la moyenne observée à une moyenne théorique dont la variance est inconnue

Pour deux échantillons indépendants

Test de Student : ttest



➤ ttest var1, by(var2) ↵

☺ Option : level, permet de fixer l'intervalle de confiance choisi

- `ttest age, by(sexe)`

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      43    10.39535     .363232    2.381871    9.662317    11.12838
       2 |      73    9.191781    .2607342    2.227714    8.672017    9.711545
---------+--------------------------------------------------------------------
combined |     116    9.637931    .2181414    2.349455    9.205835    10.07003
---------+--------------------------------------------------------------------
    diff |            1.203568     .439396                .3331279    2.074008
------------------------------------------------------------------------------
Degrees of freedom: 114

                      Ho: mean(1) - mean(2) = diff = 0

     Ha: diff < 0               Ha: diff ~= 0              Ha: diff > 0
       t =   2.7391               t =   2.7391              t =   2.7391
   P < t =   0.9964         P > |t| =   0.0072          P > t =   0.0036
```

Cette commande ne vous permet pas directement de vérifier l'hypothèse d'égalité des variances entre les groupes.

Analyse de variance à un facteur : oneway

➤ oneway var1 var2, tab ↵

```
.oneway age sexe, tab

            |           Summary of Age
       sexe |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   10.395349   2.3818714          43
          2 |   9.1917808   2.2277139          73
------------+------------------------------------
      Total |    9.637931   2.3494553         116

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      39.1989652      1   39.1989652      7.50     0.0072
 Within groups      595.594138    114   5.22450998
------------------------------------------------------------------------
    Total           634.793103    115   5.51994003

Bartlett's test for equal variances:  chi2(1) =   0.2379  Prob>chi2 = 0.626
```

↳ Ce test vous permet de vérifier l'égalité des variances entre les groupes

☺ Option : tab vous permet d'obtenir les moyennes observées entre les groupes

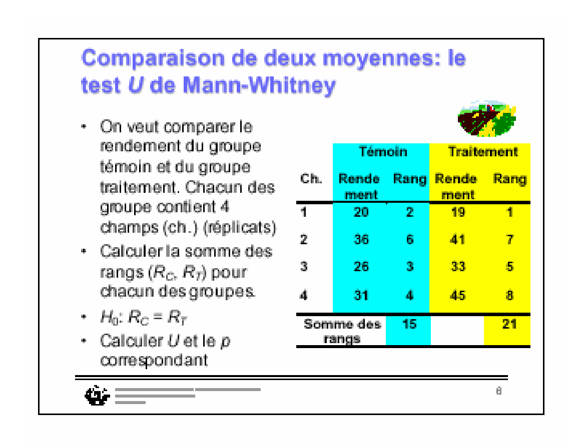
Si les conditions d'utilisation des tests ne sont pas vérifiées

↳normalité des distributions

↳variances identiques entre les groupes

Vous devez alors utiliser un test non paramétrique

Test de Mann-Whitney : ranksum



	Témoin		Traitement	
Ch.	Rendement	Rang	Rendement	Rang
1	20	2	19	1
2	36	6	41	7
3	26	3	33	5
4	31	4	45	8
Somme des rangs		15		21

➤ ranksum var1, by(var2)↵

```
.ranksum age, by(sexe)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

    sexe |      obs    rank sum    expected
---------+---------------------------------
       1 |        5        43.5          50
       2 |       14       146.5         140
---------+---------------------------------
combined |       19         190         190

unadjusted variance      116.67
adjustment for ties       -5.32
                     ----------
adjusted variance        111.35

Ho: age(sexe==1) = age(sexe==2)
             z =  -0.616
    Prob > |z| =   0.5379
```

Pour deux échantillons appariés

Ces tests sont basés sur la comparaison de la différence à 0

Test de Student pour séries appariées ttest

Condition d'utilisation : la différence a une distribution normale

➤ ttest var1 =var2↵

```
. ttest agecj5= agecj1

Paired t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  agecj5 |      98    11.39128     .213929    2.117789    10.96669    11.81586
  agecj1 |      98    10.06224      .21469    2.125322    9.636144    10.48834
---------+--------------------------------------------------------------------
    diff |      98    1.329031    .0325416    .3221454    1.264445    1.393617
------------------------------------------------------------------------------

                 Ho: mean(agecj5 - agecj1) = mean(diff) = 0

  Ha: mean(diff) < 0          Ha: mean(diff) ~= 0          Ha: mean(diff) > 0
       t =  40.8410                t =  40.8410                t =  40.8410
   P < t =   1.0000          P > |t| =   0.0000            P > t =   0.0000
```

SI : la différence est distribuée normalement
SINON utilisez un test non paramétrique

Test des rangs de Wilcoxon : signrank

➤ signrank var1=var2↵

```
.signrank  agecj5=agecj1

Wilcoxon signed-rank test

    sign |      obs   sum ranks    expected
---------+---------------------------------
positive |       16         136          68
negative |        0           0          68
    zero |        0           0           0
---------+---------------------------------
     all |       16         136         136

unadjusted variance      374.00
adjustment for ties       -0.75
adjustment for zeros       0.00
                     ----------
adjusted variance        373.25

Ho: agecj5 = agecj1
             z =   3.520
    Prob > |z| =   0.0004
```

Calculs directs sans base de données

Test d'une moyenne observée à une moyenne théorique

➢ ttesti #obs #mean #sd #val, level(#)↵

```
. ttesti 97 24 6 22

One-sample t test

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      97          24     .6092077           6    22.79073    25.20927
------------------------------------------------------------------------------
Degrees of freedom: 96

                              Ho: mean(x) = 22

     Ha: mean < 22              Ha: mean ~= 22               Ha: mean > 22
       t =   3.2830               t =   3.2830                 t =   3.2830
   P < t =   0.9993           P > |t| =   0.0014             P > t =   0.0007
```

Test de deux moyennes issues d'échantillons indépendants

➢ ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 , level(#) unequal↵

. ttesti 97 24 6 108 22 9

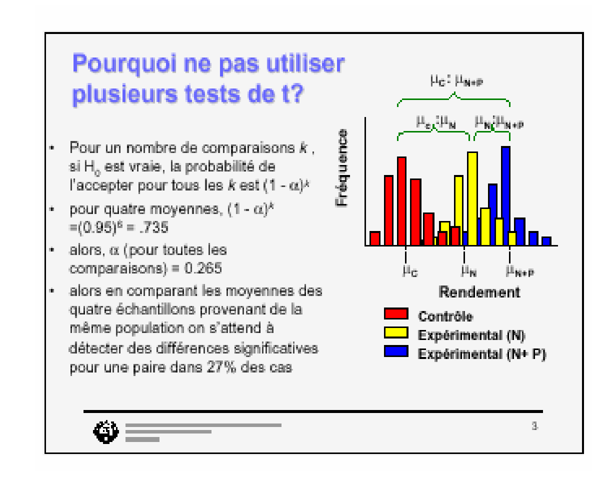
```
Two-sample t test with equal variances

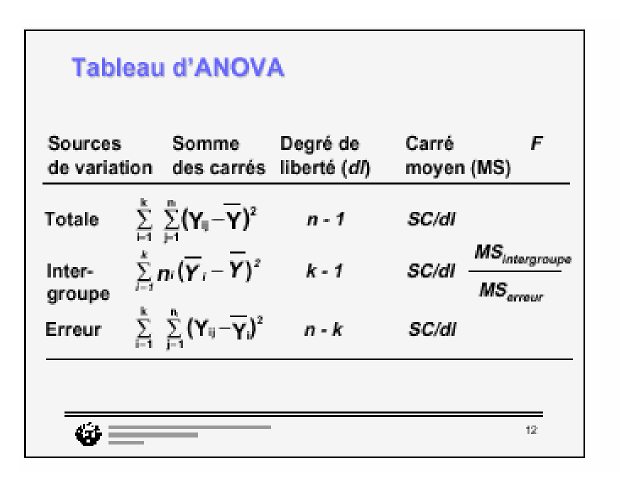
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      97          24     .6092077           6    22.79073    25.20927
       y |     108          22     .8660254           9    20.28321    23.71679
---------+--------------------------------------------------------------------
combined |     205    22.94634     .5429301    7.773576    21.87587    24.01682
---------+--------------------------------------------------------------------
    diff |                   2     1.081026               -.1314797     4.13148
------------------------------------------------------------------------------
Degrees of freedom: 203

                      Ho: mean(x) - mean(y) = diff = 0

     Ha: diff < 0               Ha: diff ~= 0                Ha: diff > 0
       t =   1.8501               t =   1.8501                 t =   1.8501
   P < t =   0.9671           P > |t| =   0.0658             P > t =   0.0329
```

Pour plus de deux échantillons indépendants : Analyse de variance





Conditions d'application de l'ANOVA

- Les résidus sont indépendants les uns des autres
- Les résidus sont distribués normalement
- La variance des résidus ne varie pas entre les traitements (homoscédasticité)
- À noter: *ces conditions s'appliquent aux résidus et non aux données brutes*
- …on doit tester les conditions d'application après que l'analyse soit faite et que les résidus soient obtenus


17

Anova

➢ Anova var1 var2⏎

```
. anova age var1

                           Number of obs =      116     R-squared     =  0.6845
                           Root MSE      = 1.33126     Adj R-squared =  0.6789

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  434.528736     2  217.264368      122.59     0.0000
                         |
                    var1 |  434.528736     2  217.264368      122.59     0.0000
                         |
                Residual |  200.264368   113  1.77225104
              -----------+----------------------------------------------------
                   Total |  634.793103   115  5.51994003
```

<u>Vérifications des conditions d'application</u>

➢ Predict res⏎
Création d'une nouvelle variable qui correspond à l'erreur d'estimation individuelle

- swilk res ↵ ou
- sktest res ↵ ou
- pnorm res ↵ ou
- kdensity res, normal

permet de tester la normalité

- Levenef res ↵ ou
- oneway res var1↵ ou
- graph res, box by(var1)

Permet de tester l'homoscédasticité

Pour pouvoir interpréter les résultats il faut faire apparaître le tableau de moyennes en fonction des groupes

- Table var1, c(mean age sd age n age)↵

```
----------+-----------------------------------
     var1 |  mean(age)      sd(age)       N(age)
----------+-----------------------------------
        1 |        4.5     .5773503            4
        2 |    8.90805     1.386246           87
        3 |         13     1.190238           25
----------+-----------------------------------
```

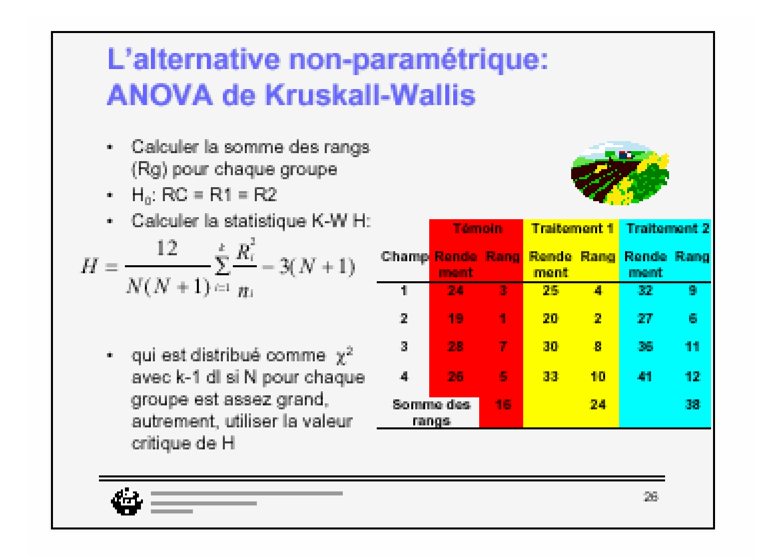
☺ Options possibles :f (%#.# f) pour formater les cellules
col pour avoir la moyenne par colonne
row pour avoir la moyenne par ligne
c pour afficher les indicateurs (maximum 5)
(min max med n mean sd)

```
table var1, c(mean age sd age n age min age max age ) f( %6.2f) row

----------+-----------------------------------------------------------
     var1 |  mean(age)      sd(age)       N(age)     min(age)     max(age)
----------+-----------------------------------------------------------
        1 |       4.50         0.58            4         4.00         5.00
        2 |       8.91         1.39           87         6.00        11.00
        3 |      13.00         1.19           25        12.00        16.00
          |
    Total |       9.64         2.35          116         4.00        16.00
----------+-----------------------------------------------------------
```

SI les conditions d'utilisations ne sont pas respectées utilisez un test non paramétrique.

Test de Kruskall et Wallis



➤ Kwallis var1, by(var2)

```
. kwallis  pj1, by(var1)

Test: Equality of populations (Kruskal-Wallis Test)

      var1           _Obs    _RankSum
         0             52     3068.50
         1             51     2945.00
         2             13      772.50

chi-squared =     0.047 with 2 d.f.
probability =     0.9766

chi-squared with ties =     0.047 with 2 d.f.
probability =     0.9766
```

Anova à deux facteurs

➤ Anova quant1 var1 var2 var1*var2

```
anova age var1 scol var1*scol

                           Number of obs =      115     R-squared     =  0.7024
                           Root MSE      = 1.31604     Adj R-squared =  0.6888

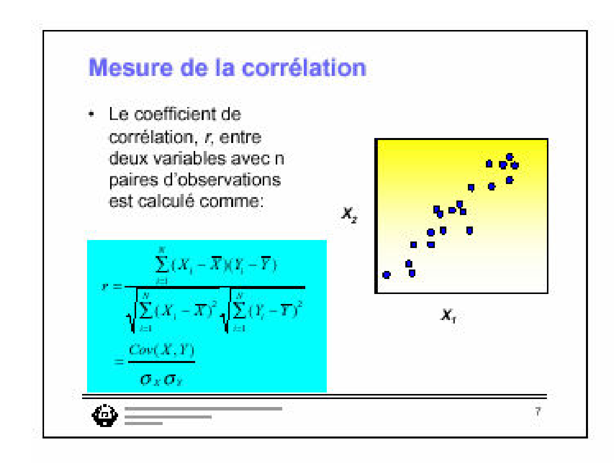
                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  445.598703     5  89.1197407      51.46     0.0000
                         |
                    var1 |  344.016225     2  172.008112      99.31     0.0000
                    scol |  2.55585275     1  2.55585275       1.48     0.2271
               var1*scol |  .370916338     2  .185458169       0.11     0.8985
                         |
                Residual |  188.783905   109  1.73196243
              -----------+----------------------------------------------------
                   Total |  634.382609   114  5.56475973
```

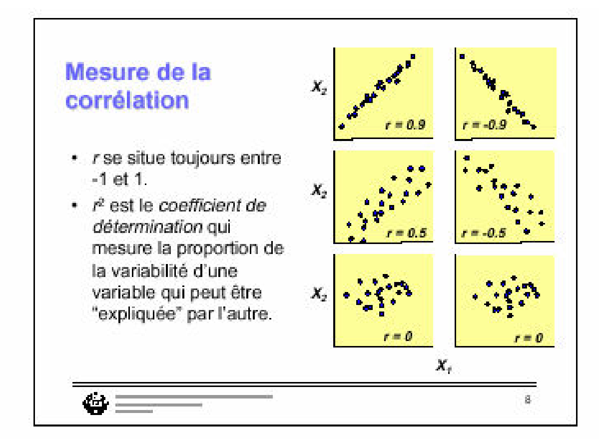
➤ Table var1 var2, c(mean quant1 sd quant1 n quant1)

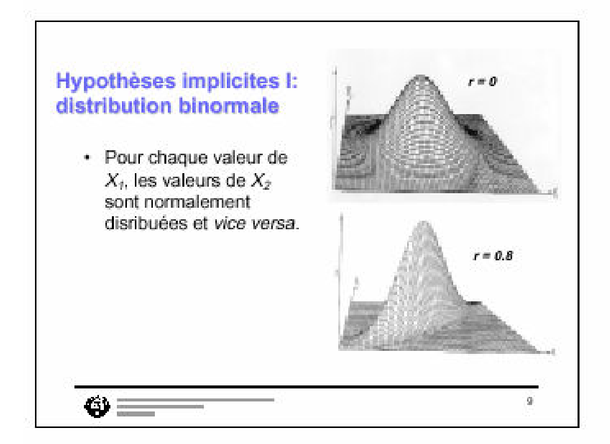
```
. table var1 scol, c(mean age sd age n age) f(%6.2f) row col

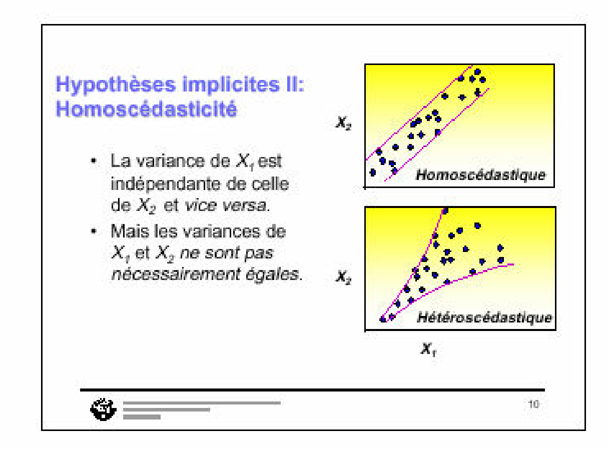
----------+--------------------
          |        scol
     var1 |     0      1  Total
----------+--------------------
        1 |  4.33   5.00   4.50
          |  0.58          0.58
          |     3      1      4
          |
        2 |  8.74   9.56   8.91
          |  1.33   1.46   1.39
          |    68     18     86
          |
        3 | 12.81  13.33  13.00
          |  1.22   1.12   1.19
          |    16      9     25
          |
    Total |  9.33  10.61   9.64
          |  2.25   2.47   2.36
          |    87     28    115
```

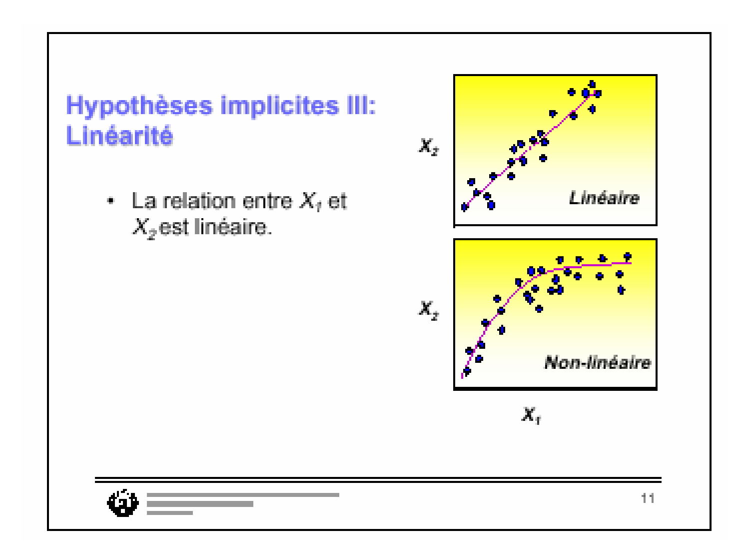
Comparaison de 2 variables quantitatives : Tests de corrélation











Coefficent de corrélation

Conditions d'application :

Régression entre x et y est linéaire
Une des 2 distributions est normale et de variance constante par rapport à l'autre

- Corr ⏎

Donne la matrice de corrélation de toutes les variables de la base
Utiliser si le nombre de variables est petit

- Corr var1 var2⏎

Donne le coefficient de corrélation de la var1 et de la var2

☺ Options possibles :

```
means donne m, sd, min et max des variables
 C donne la covariance de x et y
```

```
. corr  pbil tbil
(obs=115)

         |     pbil     tbil
---------+------------------
    pbil |   1.0000
    tbil |   0.8904   1.0000
```

```
corr  pbil tbil, c
(obs=115)

         |     pbil     tbil
---------+------------------
    pbil |   255.67
    tbil |  189.407  176.998
```

Pour tester la significativité du coefficient de corrélation vous devez utiliser un autre test

- pwcorr var1 var2⏎

☺ Options possibles :

obs : nbre de sujets
sig : test de significativité
b pour tenir compte des comparaisons multiples (bonferroni)

```
pwcorr  pbil tbil

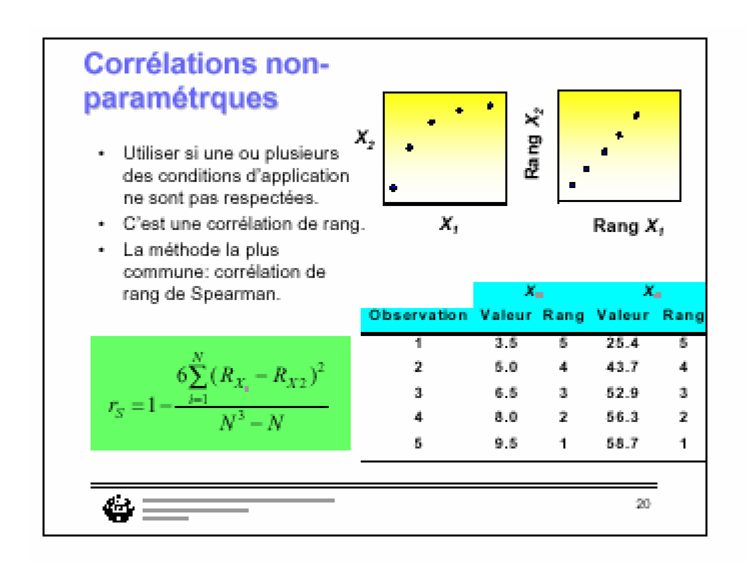
          |     pbil     tbil
----------+------------------
     pbil |   1.0000
     tbil |   0.8904   1.0000
```

```
pwcorr  pbil tbil, obs sig

          |     pbil     tbil
----------+------------------
     pbil |   1.0000
          |
          |      116
          |
     tbil |   0.8904   1.0000
          |   0.0000
          |      115      115
          |
```

Test non paramétrique de corrélation

Test de correlation des rangs de Spearman



	X		X	
Observation	Valeur	Rang	Valeur	Rang
1	3.5	5	25.4	5
2	5.0	4	43.7	4
3	6.5	3	52.9	3
4	8.0	2	56.3	2
5	9.5	1	58.7	1

H0 : les variables sont indépendantes

➢ spearman var1 var2↵

```
. spearman  pbil tbil if num<20

 Number of obs =      19
Spearman's rho =       0.8097

Test of Ho: pbil and tbil independent
      Pr > |t| =       0.0000
```

Courbes ROC (Receiver Operating Characteristic)

➢ roctab refvar var1 ↲

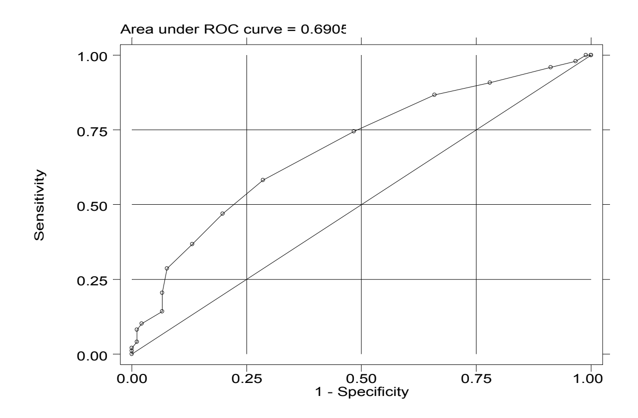
La variable de référence correspond à votre gold standard, elle doit être dichotomique codée en 0 et 1
La variable var1 est une variable quantitative continue, les valeurs les plus hautes doivent correspondre aux sujets les plus à risque

Options possibles : tab donne la table de contingence
Détail affiche la sensibilité, la spécificité, les rapports de vraisemblance positif et négatif, le % de sujets bien classés
Nograph : annule la sortie graphique
Level (#) : IC de l'aire sous la courbe
cont(#) Spécifie le nombre de groupe étudié à partir de la variable quantitative. Cette option est obligatoire si la variable quantitative possède plus de 20 valeurs distinctes.

```
. roctab volemie dppi,detail

Detailed report of Sensitivity and Specificity
------------------------------------------------------------------------------
                                           Correctly
Cut point      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= 0 )           100.00%         0.00%       51.85%       1.0000
( >= 1 )           100.00%         1.10%       52.38%       1.0111       0.0000
( >= 2 )            97.96%         3.30%       52.38%       1.0130       0.6190
( >= 3 )            95.92%         8.79%       53.97%       1.0516       0.4643
( >= 4 )            90.82%        21.98%       57.67%       1.1640       0.4179
( >= 5 )            86.73%        34.07%       61.38%       1.3155       0.3894
( >= 6 )            74.49%        51.65%       63.49%       1.5406       0.4939
( >= 7 )            58.16%        71.43%       64.55%       2.0357       0.5857
( >= 8 )            46.94%        80.22%       62.96%       2.3730       0.6614
( >= 9 )            36.73%        86.81%       60.85%       2.7857       0.7288
( >= 10 )           28.57%        92.31%       59.26%       3.7143       0.7738
( >= 11 )           20.41%        93.41%       55.56%       3.0952       0.8521
( >= 12 )           14.29%        93.41%       52.38%       2.1667       0.9176
( >= 13 )           10.20%        97.80%       52.38%       4.6429       0.9181
( >= 14 )            8.16%        98.90%       51.85%       7.4286       0.9286
( >= 15 )            4.08%        98.90%       49.74%       3.7143       0.9698
( >= 16 )            2.04%       100.00%       49.21%                    0.9796
( >= 17 )            1.02%       100.00%       48.68%                    0.9898
( >  17 )            0.00%       100.00%       48.15%                    1.0000
------------------------------------------------------------------------------


                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
         --------------------------------------------------------
           189     0.6905       0.0382        0.61570     0.76532
```

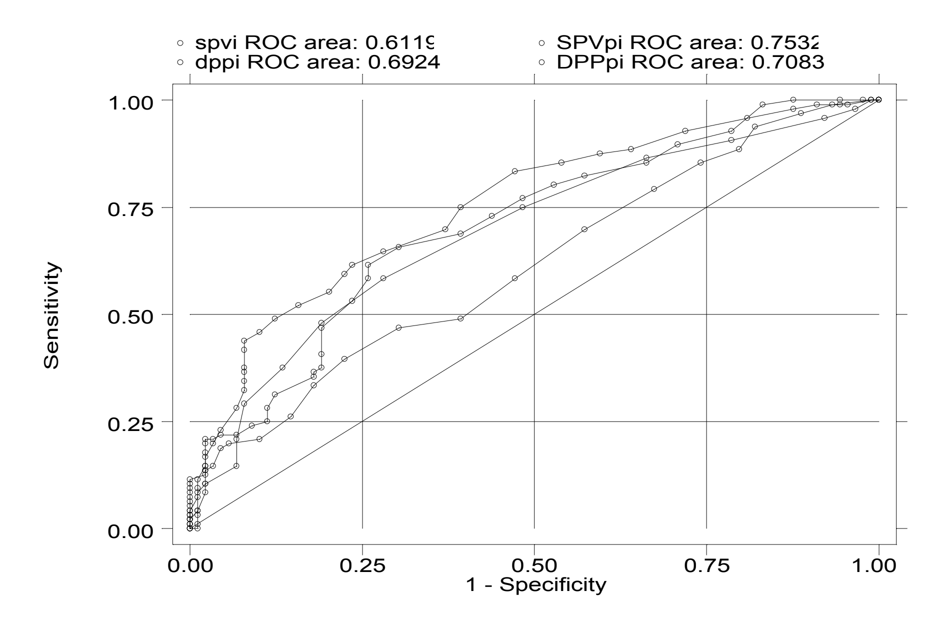


➤ Roccomp refvar var1 var2

Permet la comparaison des aires sous la courbes

```
. roccomp volemie  spvi SPVpi dppi DPPpi

                              ROC                    -Asymptotic Normal--
                   Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------------------------
DPPpi              185     0.7083       0.0378        0.63420     0.78247
SPVpi              185     0.7532       0.0354        0.68383     0.82260
dppi               185     0.6924       0.0386        0.61682     0.76801
spvi               185     0.6119       0.0411        0.53138     0.69252
-------------------------------------------------------------------------
Ho: area(DPPpi) = area(SPVpi) = area(dppi) = area(spvi)
chi2(3) =    21.40       Pr>chi2 =   0.0001
```



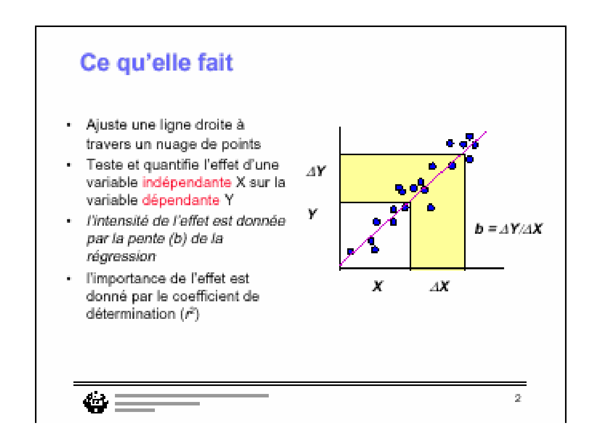
Introduction à l'analyse multivariée

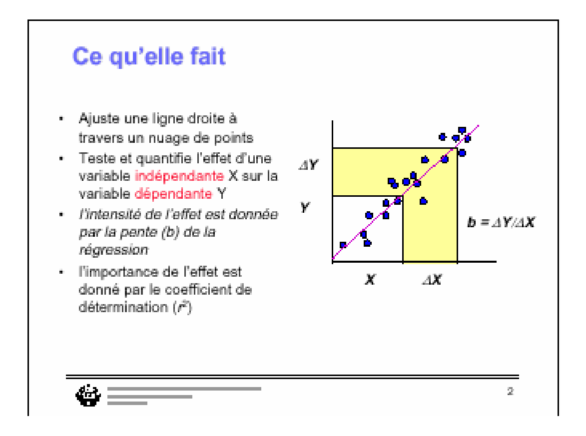
L'analyse multivariée ou modélisation vous permet de tenir compte de l'effet propre des variables que vous étudiez et des facteurs de confusion.

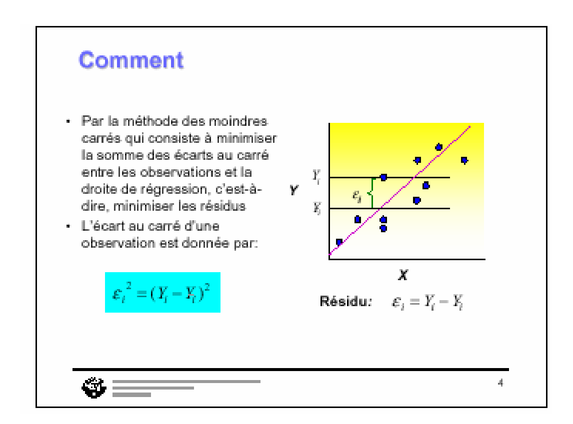
Quand la variable à expliquer est quantitative

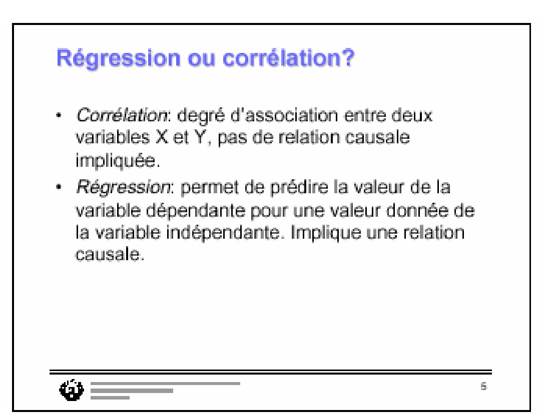
Régression linéaire

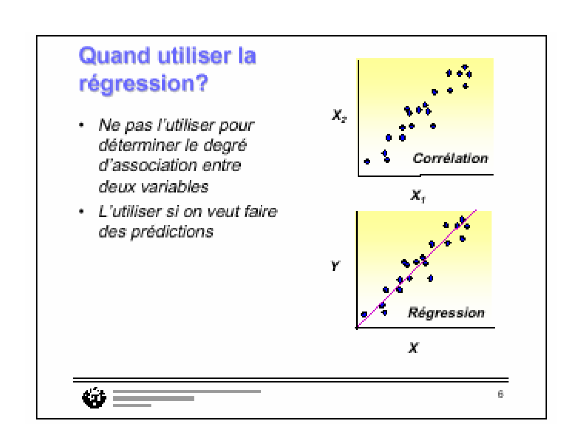
Rappel de modélisation issu du cours du Dr Morin de l'Université d'Ottawa
http://simulium.bio.uottawa.ca/bio4518/

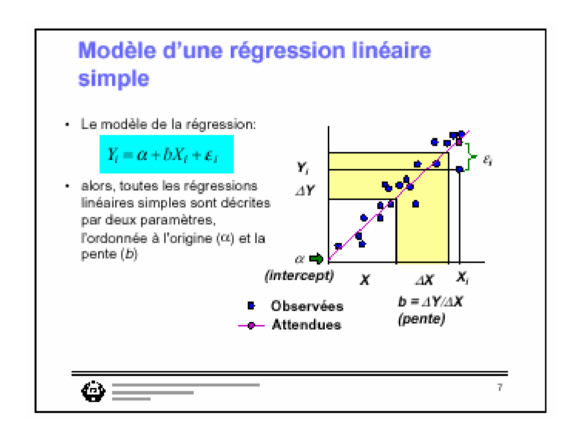


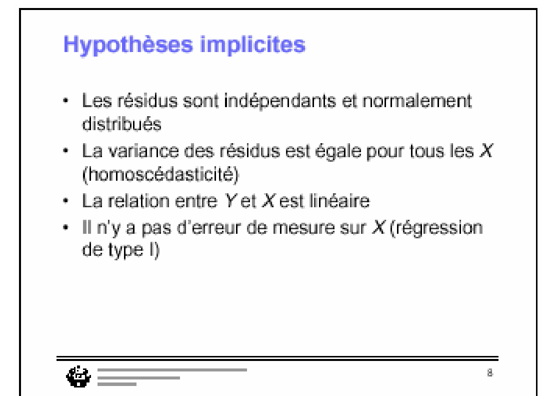


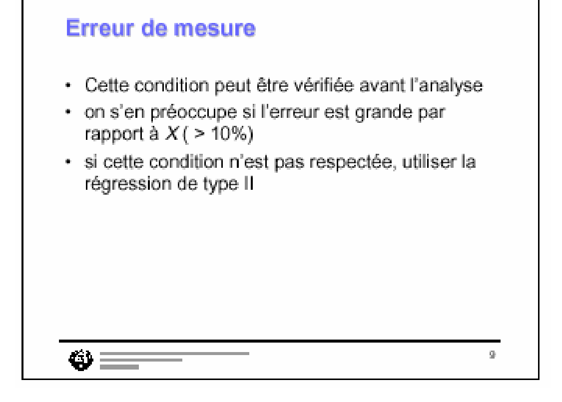


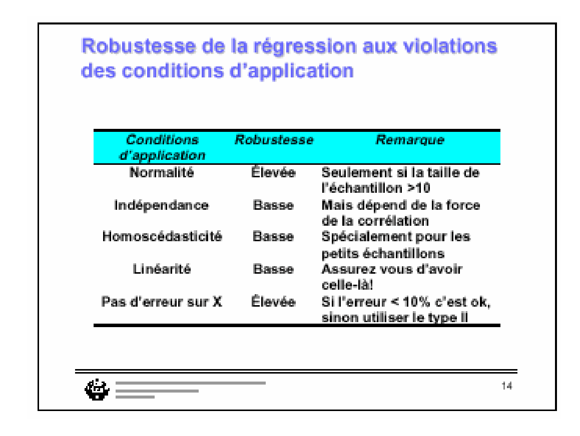








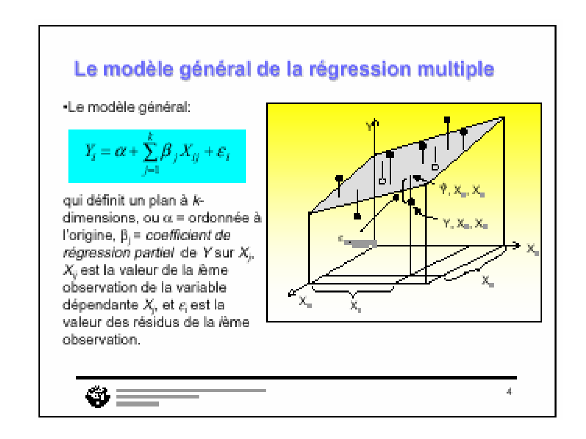




Robustesse de la régression aux violations des conditions d'application

Conditions d'application	Robustesse	Remarque
Normalité	Élevée	Seulement si la taille de l'échantillon >10
Indépendance	Basse	Mais dépend de la force de la corrélation
Homoscédasticité	Basse	Spécialement pour les petits échantillons
Linéarité	Basse	Assurez vous d'avoir celle-là!
Pas d'erreur sur X	Élevée	Si l'erreur < 10% c'est ok, sinon utiliser le type II

14



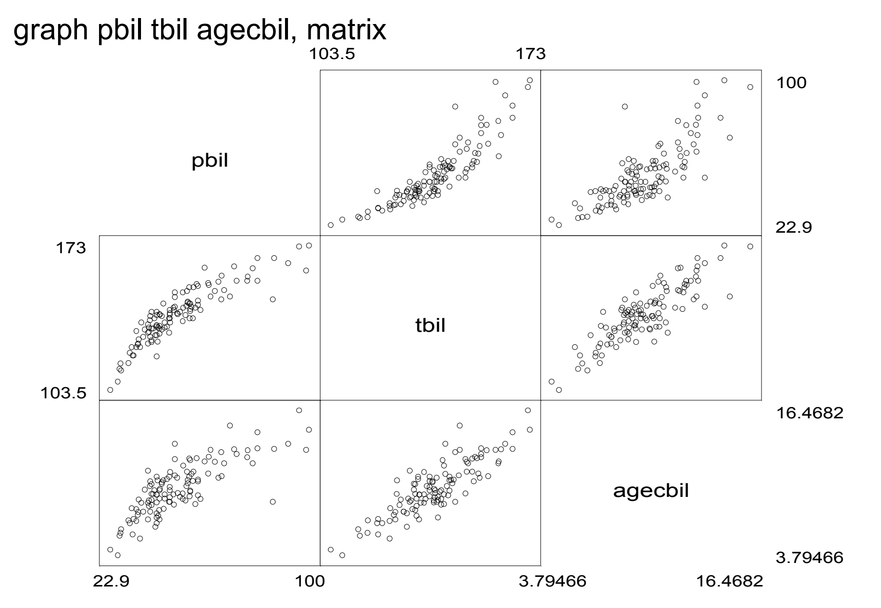
Commandes sous Stata

regress var1 var2 var3

Ex : effet de l'âge et de la taille sur le poids des enfants

Vérifier que la taille et le l'âge sont linéairement associés au poids

graph pbil tbil agecbil, matrix



Regarder la force de l'association entre le poids et l'âge et le poids et la taille

```
regress pbil agecbil

  Source |       SS       df       MS              Number of obs =     116
---------+------------------------------           F(  1,   114) =  197.47
   Model |  18500.607     1   18500.607            Prob > F      =  0.0000
Residual | 10680.2632   114  93.6865195            R-squared     =  0.6340
---------+------------------------------           Adj R-squared =  0.6308
   Total | 29180.8703   115  253.746698            Root MSE      =  9.6792

------------------------------------------------------------------------------
    pbil |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 agecbil |   5.432455   .3865821     14.053   0.000       4.666639    6.198271
   _cons |  -2.907795   3.839573     -0.757   0.450      -10.51396    4.698369


regress pbil tbil

  Source |       SS       df       MS              Number of obs =     115
---------+------------------------------           F(  1,   113) =  432.28
   Model |  23106.321     1   23106.321            Prob > F      =  0.0000
Residual | 6040.03937   113  53.4516758            R-squared     =  0.7928
---------+------------------------------           Adj R-squared =  0.7909
   Total | 29146.3604   114  255.669828            Root MSE      =  7.3111

------------------------------------------------------------------------------
    pbil |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    tbil |   1.070112   .0514688     20.791   0.000       .9681426    1.172081
   _cons |  -100.0593   7.230332    -13.839   0.000      -114.3839    -85.7347
------------------------------------------------------------------------------
```

Regarder l'effet de la taille sur le poids à âge égal

```
. regress  pbil  tbil agecbil

  Source |       SS       df       MS              Number of obs =     115
---------+------------------------------           F(  2,   112) =  218.92
   Model | 23209.4103     2  11604.7052            Prob > F      =  0.0000
Residual | 5936.95007   112  53.0084827            R-squared     =  0.7963
---------+------------------------------           Adj R-squared =  0.7927
   Total | 29146.3604   114  255.669828            Root MSE      =  7.2807

------------------------------------------------------------------------------
    pbil |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    tbil |    .949502   .1005334      9.445   0.000        .750308    1.148696
 agecbil |   .7974374   .5718241      1.395   0.166      -.3355588    1.930434
   _cons |  -90.90416   9.743854     -9.329   0.000      -110.2104   -71.59796
```

Conclusion à âge égal seule la taille explique la variation du poids

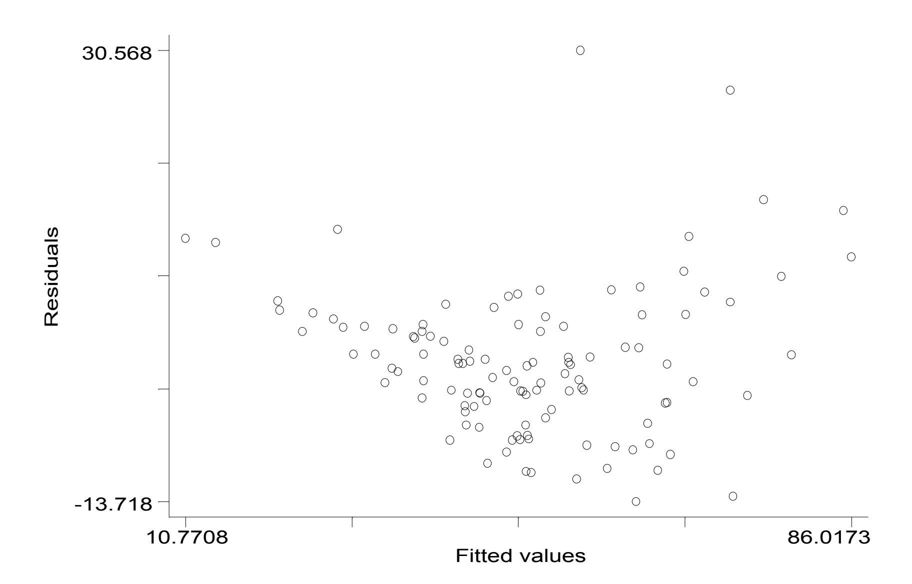
Vérification des hypothèses sous jacentes

Predict res
Calcule les résidus de la régression
swilk res ↵ ou
sktest res ↵ ou
pnorm res ↵ ou

kdensity res, normal

permet de tester la normalité

rvfplot



Permet de tester l'indépendance des résidus

Quand la variable à expliquer est dichotomique

Régression logistique

➢ **Logit** varY varX1 varX2 pour obtenir les coefficients

Où varY est la variable à expliquer et varX1, VarX2 les variables explicatives.

```
logit  varY  VarX1 VarX2 VarX3


Logit estimates                                   Number of obs   =         74
                                                  LR chi2(3)      =      29.20
                                                  Prob > chi2     =     0.0000
Log likelihood = -30.431911                       Pseudo R2       =     0.3242

------------------------------------------------------------------------------
varY           Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   VarX1 |   2.463793   .6897299     3.572   0.000       1.111947    3.815639
   VarX2 |  -.0008301   .0003781    -2.195   0.028      -.0015712   -.000089
   VarX3 |  -.9703005   .5309147    -1.828   0.068      -2.010874    .0702732
   _cons |   2.010878   1.662173     1.210   0.226      -1.246921    5.268678
------------------------------------------------------------------------------
```

➢ **Logistic** varY varX1 varX2 pour obtenir les OR

```
. logistic   varY  VarX1 VarX2 VarX3

Logit estimates                                   Number of obs   =         74
                                                  LR chi2(3)      =      29.20
```

			Prob > chi2	=	0.0000
Log likelihood = -30.431911			Pseudo R2	=	0.3242

varY	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
VarX1	11.74929	8.103839	3.572	0.000	3.040273	45.40576
VarX2	.9991702	.0003778	-2.195	0.028	.99843	.999911
VarX3	.3789691	.2012003	-1.828	0.068	.1338716	1.072801